10-29-2010 12:00 AM

Model Selection with Information Criteria

Changjiang Xu
The University of Western Ontario

Supervisor
Dr. A. Ian McLeod
The University of Western Ontario

Graduate Program in Statistics and Actuarial Sciences
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Model Selection with Information Criteria

(Spine title: Model Selection)

(Thesis format: Integrated-Article)

by

Changjiang Xu

Graduate Program
in
Statistics

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

THE UNIVERSITY OF WESTERN ONTARIO
SCHOOL OF GRADUATE AND POSTDOCTORAL STUDIES

CERTIFICATE OF EXAMINATION

Supervisor

Dr. Ian McLeod

Examiners

Dr. Duncan Murdoch

Dr. Kristina Sendova

Dr. John Knight

Dr. Paul McNicholas

The thesis by

Changjiang Xu

entitled:

Model Selection with Information Criteria

is accepted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Date

Chair of the Thesis Examination Board

ABSTRACT

This thesis is on model selection using information criteria. The information criteria include generalized information criterion and a family of Bayesian information criteria. The properties and improvement of the information criteria are investigated.

We analyze nonasymptotic and asymptotic properties of the information criteria for linear models, probabilistic models, and high dimensional models, respectively. We give probability of selecting a model and compute the probability by Monte Carlo methods. We derive the conditions under which the criteria are overfitting, consistent, or underfitting.

We further propose new model selection procedures to improve the information criteria. The procedures combine the information criteria with the probability of selecting a model and overfitting level, respectively.

In addition, we develop model selection software packages in R and examine applications to real data.

KEY WORDS: Statistical modeling, model selection, variable selection, model selection algorithm, penalized likelihood, model selection criterion, information criteria.

CO-AUTHORSHIP STATEMENT

This thesis was entirely written by Changjiang Xu under the direction of my doctoral supervisor, Dr. A. Ian McLeod.

ACKNOWLEDGEMENTS

I would like to thank my Ph.D. advisor, Professor A. Ian McLeod, for his constant support, guidance and inspiration. He always keeps his door open for us in order to discuss the research with great enthusiasm. I always felt I could approach him with absolutely any issue. I have greatly benefited both from his knowledge in a wide variety of areas and from his dedicated engagement in research. I am proud of being his student, and hope to work with him again in the future.

I would like to thank my thesis examiners, Professors Paul McNicholas, John Knight, Duncan Murdoch and Kristina Sendova for carefully reading this thesis and providing useful comments. Their comments are helpful to improve this thesis.

I would like to thank all professors in the Department of Statistical and Actuarial Sciences those who give me direct or indirect helps during my graduate study. I would like to especially thank professors Reg J. Kulperger, Serge B. Provost, Rogemar Mamon, Hao Yu, Wenqing He, Dave A. Stanford, Xiaoming Liu, and Duncan Murdoch, from those whom I took the courses and learnt a broad knowledge of statistics. I would also like to thank the department secretaries Ms. Jennifer Dungavell and Ms. Jane Bai for their kind assistances.

I would like to thank my graduate student colleagues Esam Mahdi, Juan Xiong, Lihua Yue, Nagham Mohammad, Mark Wolters, Weibin Jiang, Na Lei, Weiwei Harry Liu, Tao Jin, Mir Moosavi Avonleghi, Jing Wang, Mohammad Shahidul Islam, Radu Mitric. Finally, I would like to thank my family for their patience and love that helped me reach this point.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Statistical Modeling

There are two goals in analyzing data: extract information about the underlying system producing the data and predict the responses from future predictor variables (Breiman, 2001). Statistical modeling or data modeling is an approach toward these goals.

Statistical modeling aims at learning general rules from observed data. The data are generated as a sample from a population. Such statistical populations generating data occur widely in most areas of science including medicine, finance and engineering. From statistical point of view, the population is defined with a probability distribution.

More precisely, it is assumed that the population can be approximated by a family of probability distribution models, such as additive Gaussian models, generalized linear models, ARMA models for time series data, Weibull distribution for time-to-event data, and regression with autocorrelated or GARCH errors with financial time series. The family of distribution models may be nonparametric or mixture models, such as k-nearest-neighbor (kNN), kernel smoothing, and Bayesian networks.

The unknown distributions are then estimated from the data using some principle, such as least squares, maximum likelihood, or Bayes' rule. In general, statistical modeling involves model specification, model estimation, model selection, model validation, and model verification or adequacy checking.

1.2 Model Selection

The family of distribution models is specified to approximate the underlying system and is then estimated from the data. The next step is to assess the performance of the contending models and select the best one. The best model would have high

prediction performance, and could illustrate which predictor variables are important and how these predictors affect the response of the underlying system. The selection performance is measured by consistency, efficiency and stability. The model selection proceeds in two steps: develop an algorithm for producing contending models, also referred to as a set of candidate models; and find a criterion for ranking the contending models.

1.2.1 Algorithms

Procedures for producing the candidate models include best subsets, stepwise, or penalized methods with continuous penalty (Miller, 2002; Frank and Friedman, 1993; Tibshirani, 1996; Fan and Li, 2001; Zou, 2006; Zhang, 2010). In high dimensional statistical modeling, the traditional best subset procedure becomes infeasible due to computational cost. Furthermore, the best subset selection is unstable with a small change of data. Instead penalized maximum likelihood estimation (MLE) or least squares (LS) was suggested to automatically select significant variables with simultaneously estimating associated parameters. The attractive feature of the penalized methods is that they can produce an estimator that achieves selection consistency, stability and the oracle property. A selection procedure is said to have the oracle property if the covariance matrix of the estimates is identical to that obtained if the true model were known a priori. The key issues are the design of penalty function and algorithms for solving the penalized MLE.

1.2.2 Criteria

The model selection approaches are generally based on hypothesis testing, discrepancy, or Bayes principle (Linhart and Zucchini, 1986; McQuarrie and Tsai, 1998; Burnham and Anderson, 2002; Lahiri, 2001). The hypothesis testing based procedure usually assumes that the candidate models are nested and include a true model. The hypothesis testing with misspecified or non-nested models has also been discussed (Bera, 2000). The discrepancy is a particular class of loss functions.

Using the Bayes principle may yield two types of Bayesian selection approaches. One is the family of Bayesian information criteria, such as BIC, that is derived by approximating the posterior probability of a given model. The other is referred to as Bayesian model selection that is primarily based on the computation of the posterior

probability using Monte Carlo technique. The negative log-posterior probability may be also viewed as a loss function.

Let X and Y be the predictor and response variables. Let $\mathcal{D} = \{(x_i, y_i),\ i = 1, \ldots, n\}$ be a sample of (X, Y). Let $\hat{\mu}(X|\mathcal{D})$ be a prediction model that has been estimated from the data set $\mathcal{D}$. The errors between the response and the prediction model are measured by a loss function $L(Y, \hat{\mu}(X|\mathcal{D}))$, such as squared-error loss, log-likelihood loss, and 0-1 loss. The risk function is the expected loss function $E\{L(Y, \hat{\mu}(X|\mathcal{D}))\}$, also called expected test error (Hastie et al., 2009, §7.2). The test error is the prediction error over an independent test sample

$$E\{L(Y, \hat{\mu}(X|\mathcal{D}))|\mathcal{D}\}.$$

This expectation is taken with respect to X and Y.

Let f be the probability density function (pdf) or probability mass function (pmf) of the population generating the data. The corresponding function for the family of probability models with parameter vector θ is denoted by g_θ. The discrepancy is a functional $\Delta(g_\theta, f)$ that has the property (Linhart and Zucchini, 1986, §1.3.2)

$$\Delta(g_\theta, f) \geq \Delta(f, f).$$

The important discrepancies include Kullback-Leibler, Kolmogorov or L_∞ norm, L_1 and L_2 norms, and Pearson chi-square (Linhart and Zucchini, 1986, §2.2).

Various model selection criteria were proposed by estimating the expected loss function or discrepancy, for example,

- FPE: final prediction error, derived for linear regression models by estimating the prediction error (Akaike, 1969).
- AIC: Akaike information criterion, derived for probability models by approximating the expected Kullback-Leibler discrepancy (Akaike, 1974).
- BIC: Bayesian information criterion, obtained by approximating the negative log-posterior probability (Schwarz, 1978)
- Cross-validation and bootstrap methods: using an empirical estimate of the prediction error (Shao, 1993, 1996).

There are other model selection approaches motivated from different viewpoints, such as minimum description length (MDL) (Rissanen, 1978, 1983, 2007) and Vapnik-Chervonenkis dimension (Vapnik, 2000). The MDL approach gives a selection criterion formally identical to the family of Bayesian information criteria (Hansen and Yu, 2001).

Efficiency and Consistency

Let $\mu = E\{Y\}$. Let M_0 be the model minimizing $L(\mu, \mu(M_\alpha))$ over a set of models $\{M_\alpha\}$. Let $M_{\hat{\alpha}}$ be the model selected using a selection procedure. The selection procedure is said to be *asymptotically loss efficient* if in probability (Shao, 1997)

$$\frac{L(\mu, \mu(M_{\hat{\alpha}}))}{L(\mu, \mu(M_0))} \to 1.$$

The selection procedure is said to be *consistent* if $\Pr\{M_{\hat{\alpha}} = M_0\} \to 1$.

The M_0 is a true model or the model that is the closest to the true model. The terms overfitting and underfitting were defined two ways based on either consistency or efficiency (McQuarrie and Tsai, 1998). Using efficiency, *overfitting* is defined as choosing a model that has more variables than M_0. *Underfitting* is defined as choosing a model with too few variables compared to M_0. Under consistency, M_0 is supposed to be a true model. The procedure is *overfitting* if $M_0 \subset M_{\hat{\alpha}}$, otherwise, *underfitting* if $M_0 \subsetneq M_{\hat{\alpha}}$

1.3 Information Criteria

The FPE and AIC were respectively extended into FPE_α, the generalized FPE (Bhansali and Downham, 1977; Shibata, 1984), and AIC_α, the generalized AIC (Akaike, 1979; Bhansali, 1986). All of these criteria may be unified as a generalized information criterion (GIC)

$$\text{GIC} = -2 \log \mathcal{L}_k + \alpha c_k,$$

where $\mathcal{L}_k$ is the maximum likelihood of the model with size k, α is a positive tuning parameter that may be a constant or depend on the sample size n, and c_k reflecting the model complexity is a specified positive increasing function of the model size k. The $c_k = k$ was usually considered (Nishii, 1984; Shao, 1997).

Using Bayes or MDL principle gives a general family of Bayesian information criteria (Hansen and Yu, 2001; Rissanen, 2007)

$$\text{BIC}_\pi = -2\log \mathcal{L}_k + k\log n - 2\log \pi_k,$$

where π_k is a prior probability of the model with size k.

Information criteria include the generalized information criterion, GIC, and the family of Bayesian information criteria, BIC_π. Most of criteria derived by estimating the expected loss function or discrepancy may be considered as a special form of the GIC or BIC_π (Shao, 1997; Zhang, 2009). For example, with $\alpha = 2$ and $c_k = kn/(n-k-1)$, the GIC is the AIC_c (Hurvich and Tsai, 1989).

As compared with the bootstrap or cross-validation, the information criteria are computationally much faster which is a consideration in data mining with large datasets.

1.4 Main Issues

Asymptotic properties of the information criteria are known well (Nishii, 1984; Sin and White, 1996; Shao, 1997; Yang, 2005), but there are few results on non-asymptotic properties. The choice of the tuning parameter is a key issue for the information criteria. However, each approach to choosing the tuning parameter is identical to a model selection procedure. Also there are relatively few studies on the information criteria for high dimensional model selection (Fan and Lv, 2010). This thesis will address the three issues mentioned above. We focus on the information criteria: generalized information criterion, GIC, and the family of Bayesian information criteria, BIC_π. Summarized below are the main parts of this thesis.

Chapter 2 considers linear model selection. The properties on the GIC are analyzed and two adaptive model selection procedures are proposed.

Chapter 3 deals with the general probabilistic models. The upper bounds for the probabilities of selecting an overfitting model and the optimal model are discussed. The GIC with overfitting level is proposed.

Chapter 4 examines a family of Bayesian information criteria with the Bernoulli prior, called BIC_q. We show that the BIC_q is more effective than the usual BIC and an extended BIC.

Chapter 5 considers the high dimensional model selection. We derive the conditions under which the GIC is overfitting, consistent, or underfitting when the candidate models are estimated by the penalized maximum likelihood methods.

Chapter 2

LINEAR MODEL SELECTION USING GIC

We examine nonasymptotic and asymptotic properties of linear model selection by generalized information criterion. A necessary and sufficient condition for the model selection consistency is derived. The computation of probability of selecting a model is addressed using Monte Carlo technique and bootstrap method. Two adaptive model selection procedures are proposed based on the probability of selecting a model. These results are illustrated with simulations as well as in examples with actual data.

2.1 Introduction

By model selection we mean the choice of a best model from a set of candidate models that are often obtained by least squares or maximum likelihood estimation. The candidate models may also be provided by penalized least squares or penalized maximum likelihood, such as LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and MCP (Zhang, 2010). The true model may have infinite dimension or include unknown misspecified predictors. In this case, model selection is to find the parsimonious optimal approximation model.

Many model selection criteria have been derived based on a variety of principles such as minimizing final prediction error (Akaike, 1969, 1970), minimizing mean squared model error (Mallows, 1973), minimizing information loss (Akaike, 1974), and maximizing posterior probability (Schwarz, 1978). We mention a few of those that are most widely used. These criteria may be generalized to FPE_α (Bhansali and Downham, 1977) or AIC_α (Akaike, 1979).

For linear models, AIC_α is asymptotically equivalent to FPE_α, and both were referred to as a generalized information criterion (GIC) (Nishii, 1984; Shao, 1997). But the FPE_α is often used for the linear model. The asymptotic properties of FPE_α have

been examined by several authors (Shibata, 1984; Nishii, 1984; Shao, 1997). In practice, the sample is finite, and nonasymptotic properties are often more important than asymptotic properties.

In this chapter, we investigate the properties of FPE_α . We derive the probability distribution of selecting a model and the necessary and sufficient condition that FPE_α is consistent. The computation of the probabilities is addressed using Monte Carlo method. It is seen that the best model has the highest probability to be selected. According to this property, we propose two probability-based procedures for model selection.

The performance of FPE_α is related to a proper choice of α. How to choose the α was discussed respectively by simulation (Atkinson, 1980), approximate efficiency (Shibata, 1984), bootstrap (Rao, 1999), and generalized degrees of freedom (Shen and Ye, 2002). FPE_α using the different values of α may select the same model. So the best choice of the value of α is not unique. We derive an interval for α in which FPE_α can select a specified model. The interval is a necessary and sufficient condition under which a specified model can be selected. Each selected model corresponds to a unique interval. Thus using these interval constraint can reduce the number of the candidate models that could be selected.

A more general form of FPE_α is further considered, in which the penalty term is a monotone function of model complexity. The penalty in FPE_α includes an estimate of the variance of the model error. When the estimate is near to zero, as in high dimensional model selection, FPE_α cannot work. To avoid the drawback, we introduce a new tuning parameter that combines the unknown variance together with the parameter α, and give a modified FPE_α , denoted by FPE_γ.

2.2 Linear Model Selection

2.2.1 Optimal Model

Let $y = (y_1, ..., y_n)'$ be a vector of responses and $X = (X_1, ..., X_{d_n})$ be an $n \times d_n$ matrix, where X_k is a vector of measurements of k-th predictor. The number of predictors, d_n, may grow with the sample size n. To simplify the notation, the dependence of X, y and other random variables on n is suppressed. We mainly consider the case of deterministic predictors. When the predictors are random, the

results are still valid but some may require that $n^{-1}X(k)'X(j)$ converge almost surely for each k and j, see Assumption 2.2.

The response is expressed as $y = \mu + \varepsilon$, where $\mu = E(y|X)$ is the mean of response and $\varepsilon \sim N(0, \sigma^2 I)$. We approximate the mean μ by a linear model

$$\mu(\mathcal{S}) = X(\mathcal{S})\beta(\mathcal{S}), \tag{2.1}$$

where $\mathcal{S} = \{s_1, ..., s_k\}$ is a subset of $\{1, 2, ..., d_n\}$, $X(\mathcal{S}) = (X_{s_1}, ..., X_{s_k})$, and $\beta(\mathcal{S}) = (\beta_{s_1}, ..., \beta_{s_k})'$ is a vector of parameters that specifies the model. To simplify notation, the intercept term is considered as a predictor and included in the linear model. Each subset $\mathcal{S}$ represents a class of models. The number of components in $\mathcal{S}$, denoted by $\kappa(\mathcal{S})$, is called the model size.

By minimizing the model error

$$\text{ME}(\beta) = \| \mu - X(\mathcal{S})\beta(\mathcal{S}) \|^2, \tag{2.2}$$

the optimal linear approximation is $\mu_a = X\beta = H\mu$, where $\beta = (X'X)^- X'\mu$, $H = X(X'X)^- X'$ is a hat matrix and $(X'X)^-$ denotes the generalized inverse or Moore-Penrose pseudoinverse of $(X'X)$. The approximation error $\mu_e = (I - H)\mu$. The response may be rewritten as

$$y = \mu_a + \mu_e + \varepsilon = X\beta + \mu_e + \varepsilon.$$

Let the model class $\mathcal{S}$ be specified by

$$\beta(\mathcal{S}) = \{X(\mathcal{S})'X(\mathcal{S})\}^- X(\mathcal{S})'\mu.$$

Since μ is unknown, β is unknown but fixed for a given design matrix, and so is $\beta(\mathcal{S})$. The μ_e related to misspecified predictors is orthogonal to μ_a, and cannot be linearly predicted by X_k, $k = 1, \ldots, d_n$. The μ_a is a linear combination of the predictors.

Let $\mathcal{S}_{k_0}$ be the most parsimonious model satisfying $X(\mathcal{S}_{k_0})\beta(\mathcal{S}_{k_0}) = \mu_a$. The most parsimonious model is called as an optimal model with size $k_0 = \kappa(\mathcal{S}_{k_0})$. If the design matrix X is of full column rank, $\beta(\mathcal{S}_{k_0})$ is the nonzero elements of β. If there are no misspecified predictors, $\mu_e = 0$ and $\mathcal{S}_{k_0}$ is a true model.

Assumption 2.1. *The optimal model size, k_0, is bounded and the optimal model is identifiable, that is,*

$$\Delta = \liminf_n \min_{\substack{\mathcal{S}\neq\mathcal{S}_{k_0} \\ \kappa(\mathcal{S})\leq k_0}} n^{-1}\| X(\mathcal{S}_{k_0})\beta(\mathcal{S}_{k_0}) - X(\mathcal{S})\beta(\mathcal{S}) \| > 0.$$

Assumption 2.1 is the same as the condition used by Shao (1996, eqn. (7)). It means that the optimal model is separable from the models of size not greater than k_0. Under Assumption 2.1, the $\mathcal{S}_{k_0}$ is unique. Let $X(\mathcal{S}_{k_0}, -i)$ be the matrix $X(\mathcal{S}_{k_0})$ with the i-th column removed and $\beta(\mathcal{S}_{k_0}, -i)$ be $\beta(\mathcal{S}_{k_0})$ with the i-th element removed. Then,

$$n^{-1}\| X(\mathcal{S}_{k_0})\beta(\mathcal{S}_{k_0}) - X(\mathcal{S}_{k_0}, -i)\beta(\mathcal{S}_{k_0}, -i) \| = n^{-1}|\beta_i| \| X_i \| \geq \Delta.$$

So if the predictors in $\mathcal{S}_{k_0}$ have a finite average energy, that is, $\| X_i \|^2 /n < \infty$, then the predictors having decaying coefficients are excluded in the optimal model.

We only consider the models with size no more than $\min\{d_n, n\}$ because the model of size greater than $\min\{d_n, n\}$ would be overfitted and might not be identifiable.

2.2.2 Performance Measure

The model performance is assessed using prediction error, that is, expected squared-error loss. Let y_f be a vector of future responses at X. Thus $y_f = \mu + \varepsilon_f$, where $\varepsilon_f \sim N(0, \sigma^2 I)$. The prediction error for the model $\mathcal{S}$ is

$$\text{PE}(\hat{\beta}) = E\{\| y_f - X(\mathcal{S})\hat{\beta}(\mathcal{S}) \|^2\} = \text{ME}(\beta) + k\sigma^2 + n\sigma^2, \tag{2.3}$$

where $k = \kappa(\mathcal{S})$ and $\hat{\beta}(\mathcal{S}) = \{X(\mathcal{S})'X(\mathcal{S})\}^- X(\mathcal{S})'y$. The estimated model error is

$$\text{ME}(\hat{\beta}) = E\{\| \mu - X(\mathcal{S})\hat{\beta}(\mathcal{S}) \|^2\} = \text{ME}(\beta) + k\sigma^2.$$

Let $\text{RSS}(\hat{\beta}) = \| y - X(\mathcal{S})\hat{\beta}(\mathcal{S}) \|^2$ be the residual sum of squares (RSS). Then

$$E\{\text{RSS}(\hat{\beta})\} = \mu'(I - H)\mu + \sigma^2(n - k) = \text{ME}(\beta) + \sigma^2(n - k).$$

Thus

$$\text{PE}(\hat{\beta}) = \text{ME}(\hat{\beta}) + n\sigma^2 = E\{\text{RSS}(\hat{\beta})\} + 2k\sigma^2. \tag{2.4}$$

Minimizing $\text{PE}(\hat{\beta})$ is not equivalent to minimizing $\text{ME}(\beta)$. Usually there is a trade-off between the model size and the model error to minimize the prediction error. The following proposition elucidates this relationship. Let $k = \kappa(\mathcal{S})$ and

$$\text{SNR} = \min_{k<k_0} \frac{\parallel X(\mathcal{S}_{k_0})\beta(\mathcal{S}_{k_0}) - X(\mathcal{S})\beta(\mathcal{S}) \parallel^2}{(k_0 - k)\sigma^2},$$

be an average signal-to-noise ratio.

Proposition 2.1. *Let $\mathcal{S}_{k_{\text{PE}}}$ be the model that minimizes prediction error. Under Assumption 2.1, if* SNR > 1, *$\mathcal{S}_{k_{\text{PE}}} = \mathcal{S}_{k_0}$, otherwise, if* SNR < 1, *$k_{\text{PE}} < k_0$.*

The proof of Proposition 2.1 and all others are in Appendix 2.9. If SNR $= 1$, then either $k_{\text{PE}} < k_0$ or $\mathcal{S}_{k_{\text{PE}}} = \mathcal{S}_{k_0}$. The SNR is related to the sample size and the noise level. As the sample size increases, the SNR becomes large and the model having a good prediction tends to the optimal model. That is, the optimal model $\mathcal{S}_{k_0}$ might minimize the prediction error. But the $\mathcal{S}_{k_0}$ is unknown and needs to be estimated. Selecting a model is equivalent to estimating $\mathcal{S}_{k_0}$.

2.2.3 Model Selection

We consider the problem of selecting the best model from a set of candidate models denoted by $\{\mathcal{S}_k, k = 1, \ldots, K\}$. The best model is selected by the minimum value of some selection criterion. Each candidate model $\mathcal{S}_k$ is the best model in the set of models with size k, which has the minimum RSS. Often stepwise methods or the branch-and-bound algorithm (Furnival and Wilson, 1974; Gatu, 2006) may be used but other optimization methods are also available for larger space problems (Hofmann et al., 2007). Penalized least squares, such as LASSO and SCAD, can also be employed to get the candidate models.

Assume that there is no multicollinearity for each model $\mathcal{S}_k$. Then

$$\hat{\beta}(\mathcal{S}_k) = \{X(\mathcal{S}_k)'X(\mathcal{S}_k)\}^{-1}X(\mathcal{S}_k)'y.$$

Assume that the sizes of the candidate models are unique, that is, each size corresponds to one candidate model. If the model sizes are not unique, we keep one model

for each size that has a minimum RSS. Hence selecting a model is equivalent to selecting the model size. For simplicity of notation, let $X(k) = X(\mathcal{S}_k)$ and $\hat{\beta}(k) = \hat{\beta}(\mathcal{S}_k)$ in the sequel.

Let $\mathcal{S}_{k_n}$ be the selected model with size k_n. If $\Pr\{k_n = k_0\} \to 1$ as $n \to \infty$, the model selection procedure is consistent. If $k_n > k_0$, the selected model is overfitting. Otherwise, if $k_n < k_0$, the selected model is underfitting.

2.2.4 Selection Criterion

Most model selection criteria proposed to estimate the prediction error can be unified into a generalized information criterion (GIC). For a linear model, the GIC has the form,

$$\text{GIC} = n \log \hat{\sigma}_k^2 + \alpha k,$$

where α is a tuning parameter, $\hat{\sigma}_k^2 = \text{RSS}_k/n$, and $\text{RSS}_k = \| y - X(k)\hat{\beta}(k) \|^2$. The GIC is asymptotically equivalent to a generalized FPE,

$$\text{FPE}_\alpha = n\hat{\sigma}_k^2 + \alpha k s_K^2 = \text{RSS}_k + \alpha k s_K^2, \tag{2.5}$$

where $s_k^2 = \text{RSS}_k/(n-k)$, since as $\alpha k/n \to 0$,

$$\frac{\log\{\text{FPE}_\alpha/n\}}{\text{GIC}/n} = \frac{\log \hat{\sigma}_k^2 + \log\{1 + (\alpha k/n)(s_K^2/\hat{\sigma}_k^2)\}}{\log \hat{\sigma}_k^2 + \alpha k/n} \xrightarrow{a.s.} 1.$$

So, for the linear model selection, we focus on the FPE_α instead of the GIC. Using FPE_α, we may get the finite-sample distribution of selecting a model instead of an asymptotic distribution.

The original FPE proposed by Akaike (1969, 1970) is $\text{FPE} = \{1 + 2k/(n-k)\}\hat{\sigma}_k^2$. Bhansali and Downham (1977) generalized the FPE into $\text{FPE}^B = (1 + \alpha k/n)\hat{\sigma}_k^2$. Shibata (1984) suggested the FPE_α. Both FPE_α and FPE_α^B are asymptotically equivalent since as $\alpha k/n \to 0$,

$$\frac{\text{FPE}_\alpha}{n\text{FPE}_\alpha^B} = \frac{1 + (\alpha k/n)(s_K^2/\hat{\sigma}_k^2)}{1 + \alpha k/n} \xrightarrow{a.s.} 1.$$

The expected FPE_α is

$$E\{\text{FPE}_\alpha\} = \text{ME}(\mathcal{S}_k) + k\sigma^2(\alpha_\delta - 1) + n\sigma^2, \tag{2.6}$$

where $\alpha_\delta = \alpha(1+\delta_n^2/\sigma^2)$, $\delta_n^2 = \mu_e'\mu_e/(n-K)$. If there are no misspecified predictors, $\mu_e = 0$ and $\alpha_\delta = \alpha$. If $\alpha_\delta = 1$, the expected FPE_α equals the quadratic risk. If $\alpha_\delta = 2$, it equals the prediction error (2.3). Comparing (2.6) with (2.3), from Proposition 2.1 we have

Proposition 2.2. *Assume $\mathcal{S}_{k_0} \in \{\mathcal{S}_k\}$. Let $\mathcal{S}_{k_{\text{FPE}_\alpha}}$ be the model minimizing the expected FPE_α. Under Assumption 2.1, if $0 < \alpha_\delta - 1 < \text{SNR}$ then $\mathcal{S}_{k_{\text{FPE}_\alpha}} = \mathcal{S}_{k_0}$, otherwise, if $\alpha_\delta - 1 > \text{SNR}$ then $k_{\text{FPE}_\alpha} < k_0$. Furthermore, if $1 \le \alpha_\delta - 1 < \text{SNR}$, $\mathcal{S}_{k_{\text{FPE}_\alpha}} = \mathcal{S}_{k_0} = \mathcal{S}_{k_{\text{PE}}}$.*

Proposition 2.2 shows that the model minimizing the expected FPE_α may have the minimum model error and prediction error if the sample size is large enough or equivalently the SNR is higher.

2.3 Nonasymptotic Properties

We analyze the probability of selecting a model by FPE_α and the computation of the probability using the Monte Carlo method. The following lemma provides the interval of α, in which FPE_α can select a specified model.

Lemma 2.1. FPE_α *can select the model $\mathcal{S}_k$ if and only if*

$$\max_{j>k} A_{k,j} \le \alpha \le \min_{j<k} A_{k,j},$$

where $A_{k,j} = (\text{RSS}_k - \text{RSS}_j)/\{(j-k)s_K^2\}$ for $j \ne k$, $k = 1, \ldots, K$. Let k_i, $i = 1, ..., m$, be all of sizes selected by FPE_α with different $\alpha = \alpha_i$ and be in ascending order. Then $A_{k_1,k_2} \le \alpha_1 < \infty$,

$$A_{k_i,k_{i+1}} \le \alpha_i \le A_{k_{i-1},k_i},$$

and $0 \le \alpha_m \le A_{k_{m-1},k_m}$. If $\alpha = A_{k_i,k_{i+1}}$, the FPE_α may select $\mathcal{S}_{k_i}$ and $\mathcal{S}_{k_{i+1}}$.

Here we define $\min_{j<1} A_{1,j} = \infty$ and $\max_{j>K} A_{K,j} = 0$. Let $\kappa(\alpha)$ be the model size selected by FPE_α. From Lemma 2.1, the probability of selecting the model $\mathcal{S}_k$ is

$$\Pr\{\kappa(\alpha) = k\} = \Pr\{\max_{j>k} A_{k,j} \le \alpha \le \min_{j<k} A_{k,j}\}. \tag{2.7}$$

Define an indicator function $I(X, y, \alpha) = I\{\max_{j>k} A_{k,j} \leq \alpha \leq \min_{j<k} A_{k,j}\}$. Then

$$\Pr\{\kappa(\alpha) = k\} = E\{I(X, y, \alpha)\}.$$

Let $X^{(i)}$ and $y^{(i)}$ be the sample of X and y. By the strong law of large numbers,

$$p_k(\alpha) = \frac{1}{N} \sum_{i=1}^{N} I(X^{(i)}, y^{(i)}, \alpha) \xrightarrow{a.s.} \Pr\{\kappa(\alpha) = k\}. \tag{2.8}$$

The probability, $\Pr\{\kappa(\alpha) = k\}$, can also be estimated using the bootstrap method or resampling the data $\{X,\ y\}$.

2.3.1 Nested Subsets

Assume that the candidate models are nested. In order to compute the probability in (2.7), we analyze the distribution of $A_{k,j}$. The following Lemma 2.2 holds from the distribution of quadratic forms (Rao, 1973, §3b.4).

Lemma 2.2. *Let $\mathcal{S}_j \supset \mathcal{S}_k \supset \mathcal{S}_l$. Let $\chi^2_d(\lambda)$ denote a noncentral chi-square distribution with a degree of freedom d and a noncentrality parameter λ. Then*

$$\begin{aligned} \mathrm{RSS}_k/\sigma^2 &= y'(I - H_k)y/\sigma^2 \sim \chi^2_{n-k}(\lambda_k), \\ (\mathrm{RSS}_k - \mathrm{RSS}_j)/\sigma^2 &= y'(H_j - H_k)y/\sigma^2 \sim \chi^2_{j-k}(\lambda_{kj}), \end{aligned}$$

where $\sigma^2\lambda_k = \mu'(I - H_k)\mu$ and $\sigma^2\lambda_{kj} = \mu'(H_j - H_k)\mu$. If $\mathcal{S}_k \supset \mathcal{S}_{k_0}$, $\lambda_{k_0,k} = 0$. Furthermore, RSS_K, $\mathrm{RSS}_l - \mathrm{RSS}_k$ and $\mathrm{RSS}_k - \mathrm{RSS}_j$ are independent.

Since $\{\mathcal{S}_k\}$ are nested, from Lemma 2.2, for $k = 1, \ldots, K-1$,

$$Z_k = (\mathrm{RSS}_k - \mathrm{RSS}_{k+1})/\sigma^2 \sim \chi^2_1(\lambda_{k,k+1}),$$

are independent, where $\lambda_{k,k+1} = E\{\mathrm{RSS}_k - \mathrm{RSS}_{k+1}\}/\sigma^2 - 1 = \mu'(H_{k+1} - H_k)\mu/\sigma^2$ are the noncentrality parameters. Let $Z_K = s^2_K/\sigma^2$. Then for $j = 1, \ldots, k-1$,

$$A_{k,j} = (Z_j + \cdots + Z_{k-1})/Z_K(k - j),$$

and for $j = k+1, \ldots, K$,

$$A_{k,j} = (Z_k + \cdots + Z_{j-1})/Z_K(j-k).$$

Let $\mathbf{A}_k = (A_{k,1}, \ldots, A_{k,k-1}, A_{k,k+1}, \ldots, A_{k,K})$ and $\mathbf{Z} = (Z_1, \ldots, Z_{K-1})$ be vectors of length $K-1$, and for $1 < k < K$,

$$\mathbf{G}_k = \begin{bmatrix} \frac{1}{k-1} & \frac{1}{k-1} & \cdots & \frac{1}{k-1} & & & & \\ & \frac{1}{k-2} & \cdots & \frac{1}{k-2} & & \mathbf{0} & & \\ & & \ddots & \vdots & & & & \\ & & & 1 & & & & \\ & & & & 1 & & & \\ & & & & \frac{1}{2} & \frac{1}{2} & & \\ & \mathbf{0} & & & \vdots & \vdots & \ddots & \\ & & & & \frac{1}{K-k} & \frac{1}{K-k} & \cdots & \frac{1}{K-k} \end{bmatrix}$$

Then

$$\mathbf{A}_k = \mathbf{G}_k \mathbf{Z}/Z_K. \tag{2.9}$$

Define an indicator function $I(\mathbf{A}_k, \alpha) = I\{\max_{j>k} A_{k,j} \le \alpha \le \min_{j<k} A_{k,j}\}$. From (2.7), for $1 < k < K$,

$$\Pr\{\kappa(\alpha) = k\} = E\{I(\mathbf{A}_k, \alpha)\} = E\{I(\mathbf{G}_k\mathbf{Z}/Z_K, \alpha)\}.$$

From Lemma 2.2, Z_k, $k = 1, \ldots, K$, are independent. Assume that there are no misspecified predictors, that is, $\mu_e = 0$. Then $Z_K = s_K^2/\sigma^2 \sim \chi^2_{n-K}/(n-K)$. Let $Z_k^{(i)}$, $i = 1, \ldots, N$, be *i.i.d.* sample of Z_k. Let $\mathbf{Z}^{(i)} = (Z_1^{(i)}, \ldots, Z_{K-1}^{(i)})$. By strong law of large numbers,

$$p_k^{Nest}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} I(\mathbf{G}_k\mathbf{Z}^{(i)}/Z_K^{(i)}, \alpha) \xrightarrow{a.s.} \Pr\{\kappa(\alpha) = k\}. \tag{2.10}$$

Sampling Z_k needs the noncentrality parameter $\lambda_{k,k+1} = E\{\text{RSS}_k - \text{RSS}_{k+1}\}/\sigma^2 - 1$. An unbiased estimator of the non-centrality parameter is, see Kubokawa et al.

(1993),

$$\hat{\lambda}_{k,k+1} = \max\{\frac{(n-K-2)(\text{RSS}_k - \text{RSS}_{k+1})}{(n-K)s_K^2} - 1, 0\}. \tag{2.11}$$

Using the estimates of noncentrality parameters yields a conditional probability. If the estimate is consistent, the conditional probability will converge to the true probability.

2.3.2 Unnested Subsets

Let X be the $n \times K$ design matrix corresponding to the subsets $\{\mathcal{S}_k\}$. Consider the matrix decomposition $X = UV$, where U is an $n \times K$ column orthogonal matrix, that is, $U'U = I_K$, and V is a $K \times K$ square matrix. This decomposition can be constructed by the singular value decomposition. Then $X(k) = UV(k)$, where $V(k) = V(\mathcal{S}_k)$.

Let $\widetilde{H}_k = V(k)\{V(k)'V(k)\}^{-1}V(k)'$, $\tilde{\mu} = U'\mu$, $\tilde{\varepsilon} = U'\varepsilon$, and $\epsilon = \tilde{\varepsilon}/\sigma$. Then $y'H_k y = (\tilde{\mu} + \tilde{\varepsilon})'\widetilde{H}_k(\tilde{\mu} + \tilde{\varepsilon})$, and

$$(j-k)A_{k,j} = y'(H_j - H_k)y/s_K^2 = (\tilde{\mu}/\sigma + \epsilon)'(\widetilde{H}_j - \widetilde{H}_k)(\tilde{\mu}/\sigma + \epsilon)/Z_K.$$

The indicator function $I(\mathbf{A}_k, \alpha)$ can be rewritten as

$$I(\mathbf{A}_k, \alpha) = I(\epsilon, Z_K, \mu, \sigma, \alpha). \tag{2.12}$$

Since $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$, $\epsilon \sim N(\mathbf{0}, I_K)$. Assume that there are no misspecified predictors, that is, $\mu_e = 0$. Then $Z_K \sim \chi^2_{n-K}/(n-K)$. Let $\epsilon^{(i)}$ and $Z_K^{(i)}$, $i = 1, \ldots, N$, be *i.i.d.* sample of ϵ and Z_K, respectively. By strong law of large numbers,

$$p_k^{Unnest}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} I(\epsilon^{(i)}, Z_K^{(i)}, \mu, \sigma, \alpha) \xrightarrow{a.s.} \Pr\{\kappa(\alpha) = k\}. \tag{2.13}$$

The indicator function $I(\epsilon, Z_K, \mu, \sigma, \alpha)$ includes the unknown parameters μ and σ. In practice, we instead use the estimates $\hat{\sigma}^2 = s_K^2$ and $\hat{\mu} = Hy$.

2.3.3 Upper Bounds

Let $\widetilde{\mathcal{S}}_{k-1}$ and $\widetilde{\mathcal{S}}_{k+1}$ be the models of size $k-1$ and $k+1$, respectively, and satisfy $\widetilde{\mathcal{S}}_{k-1} \subset \mathcal{S}_k \subset \widetilde{\mathcal{S}}_{k+1}$. If the candidate models are nested, $\widetilde{\mathcal{S}}_{k-1} = \mathcal{S}_{k-1}$ and $\widetilde{\mathcal{S}}_k = \mathcal{S}_k$.

Since $\mathcal{S}_j$ has the minimum RSS in the models of size j, $\widetilde{\text{RSS}}_j \geq \text{RSS}_j$ for $j = k-1$ and $k+1$. Assume that there are no misspecified predictors, that is, $\mu_e = 0$. Then $Z_K \sim \chi^2_{n-K}/(n-K)$.

From (2.7) and Lemma 2.2, an upper bound of the probability that FPE_α selects the model $\mathcal{S}_k$, $1 < k < K$, is

$$\begin{aligned} \Pr\{\kappa(\alpha) = k\} &\leq \Pr\{A_{k,k+1} \leq \alpha \leq A_{k,k-1}\} \\ &= \Pr\{(\text{RSS}_k - \text{RSS}_{k+1})/s_K^2 \leq \alpha \leq (\text{RSS}_{k-1} - \text{RSS}_k)/s_K^2\} \\ &\leq \Pr\{(\text{RSS}_k - \widetilde{\text{RSS}}_{k+1})/s_K^2 \leq \alpha \leq (\widetilde{\text{RSS}}_{k-1} - \text{RSS}_k)/s_K^2\} \\ &= \Pr\{F_{1,n-K}(\tilde{\lambda}_{k,k+1}) \leq \alpha\} \Pr\{\alpha \leq F_{1,n-K}(\tilde{\lambda}_{k-1,k})\}, \end{aligned} \tag{2.14}$$

where $\tilde{\lambda}_{k,k+1}$ and $\tilde{\lambda}_{k-1,k}$ are noncentrality parameters of $(\text{RSS}_k - \widetilde{\text{RSS}}_{k+1})/\sigma^2$ and $(\widetilde{\text{RSS}}_{k-1} - \text{RSS}_k)/\sigma^2$, respectively, and $F_{d_1,d_2}(\lambda)$ represents a noncentral F-distribution with degrees of freedom d_1 and d_2 and noncentrality parameter λ.

If $k > k_0$, $\tilde{\lambda}_{k,k+1} = 0$ and $\tilde{\lambda}_{k-1,k} = 0$. Hence, the probability of selecting a overfitted model by FPE_α is bounded by,

$$\Pr\{\kappa(\alpha) = k\} \leq \Pr\{F_{1,n-K} \leq \alpha\}(1 - \Pr\{F_{1,n-K} \leq \alpha\}) \triangleq p. \tag{2.15}$$

Similarly, the probability of selecting the true model by FPE_α is bounded by

$$\Pr\{\kappa(\alpha) = k_0\} \leq \Pr\{F_{1,n-K} \leq \alpha\}(1 - \Pr\{F_{1,n-K}(\lambda) \leq \alpha\}) \triangleq p_0, \tag{2.16}$$

where $\lambda = \tilde{\lambda}_{k_0-1,k_0}$ is a noncentrality parameter of $(\widetilde{\text{RSS}}_{k_0-1} - \text{RSS}_{k_0})/\sigma^2$.

The upper bound probabilities of selecting an overfitted model and the true model are plotted in Figure 2.1. Each curve in the right of Figure 2.1 corresponds to a different non-centrality parameter λ. The small λ represents the case of small sample size. The circles represent the maximum values of each curve.

When K is large, the following upper bounds may also be used to reduce the computational burden.

$$\Pr\{\kappa(\alpha) = k\} \leq \Pr\{\max_{k<j<k+h} A_{k,j} \leq \alpha \leq \min_{k-h<j<k} A_{k,j}\},$$

where $0 < h < K$ is a specified number, for example, $h = 4$.

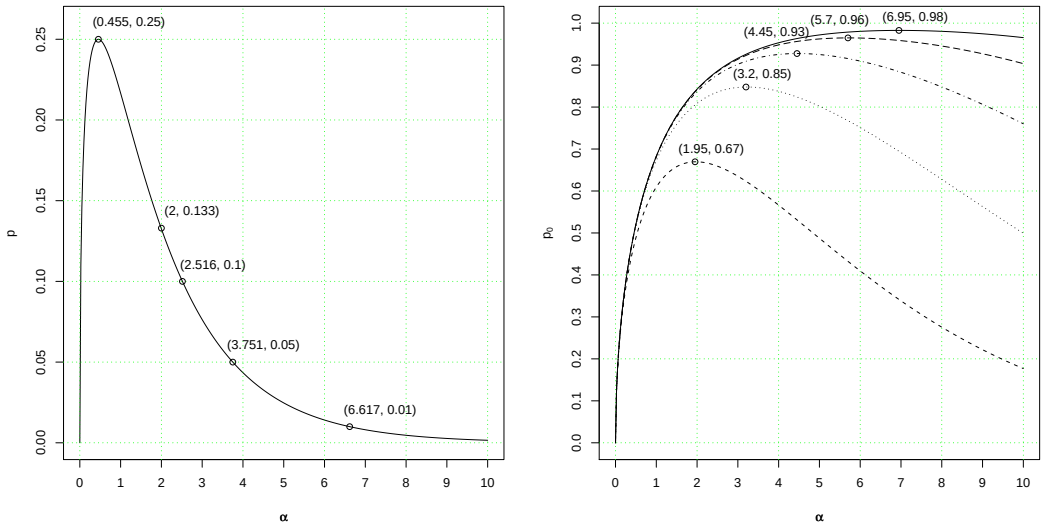

Figure 2.1: Left: upper bound of the probability of selecting a overfitted model. Right: upper bound of the probability of selecting the true model. The non-centrality parameters, λ, are 5 (dashed), 10 (dotted), 15 (dotdash), 20 (longdash), and 25 (solid), respectively.

2.4 Asymptotic Properties

For convenience, denote $\Pr\{A_n\} \overset{asy.}{=} \Pr\{B_n\}$ if $\Pr\{\lim A_n\} = \Pr\{\lim B_n\}$. To analyze the asymptotic properties on FPE_α, we use the following assumption.

Assumption 2.2. *There almost surely exist the limits:* $n^{-1}y'y \xrightarrow{a.s.} v^2$, $n^{-1}X(k)'y \xrightarrow{a.s.} v_k$, *and* $n^{-1}X(k)'X(j) \xrightarrow{a.s.} V_{kj}$.

Lemma 2.3. *Under Assumptions 2.1 and 2.2,*

$$\begin{aligned} n^{-1}\text{ME}(\beta(k)) &\xrightarrow{a.s.} \Delta_k^2 + \delta^2, \\ n^{-1}\text{RSS}(\beta(k)) &\xrightarrow{a.s.} \Delta_k^2 + \delta^2 + \sigma^2, \end{aligned}$$

where

$$\Delta_k^2 = \lim n^{-1} \| X(k_0)\beta(k_0) - X(k)\beta(k) \|^2 = v'_{k_0} V_{k_0 k_0}^{-1} v_{k_0} - v'_k V_{kk}^{-1} v_k,$$

$$\delta^2 = \lim n^{-1} \mu'_e \mu_e = v^2 - \sigma^2 - v'_{k_0} V_{k_0 k_0}^{-1} v_{k_0}.$$

Moreover, $\Delta_k^2 = 0$ *if and only if* $k \geq k_0$. *If* $k < k_0$, $\Delta_k^2 \geq \Delta$.

Proposition 2.3. *Assume that Assumptions 2.1 and 2.2 hold. Let $\mathcal{S}_k$ be the best model of size k, $k = 1, \ldots, K$, that minimizes* RSS. *If the sample size is sufficiently large, then $\mathcal{S}_k$ is independent of the sample for a fixed k, $\mathcal{S}_{k_0} \in \{\mathcal{S}_k\}$, and either $\mathcal{S}_k \subset \mathcal{S}_j$ or $X(k)$ and $X(j)$ span the same subspace for the fixed k and j with $k < j$.*

It may not hold that $\mathcal{S}_k \supset \mathcal{S}_{k_0}$ for $k > k_0$ if the sample size is small. For example, there are three predictors x_1, x_2 and $x_3 = x_1 + x_2 + 0.01e$, and the true model $y = x_3 + 0.1e$, where $e \sim N(0, 1)$. Let $n = 20$. The first two best subsets may be $\mathcal{S}_1 = \{x_3\}$ and $\mathcal{S}_2 = \{x_1, x_2\}$.

From Lemma 2.3, there exists the following limit

$$\tau = \lim_{n\to\infty} \min_{j<k_0} n^{-1} A_{k_0,j} \overset{a.s.}{=} \min_{j<k_0} \frac{\Delta_j^2}{(\sigma^2 + \delta^2)(k_0 - j)} > 0. \tag{2.17}$$

Theorem 2.1. *Let $\alpha = \alpha_n$ and $n^{-1}\alpha_n \to r$. Let $\kappa(\alpha)$ be the model size selected by* FPE_α. *Assumptions 2.1 and 2.2 hold.*

1) If $\alpha < \infty$, then for $k_0 < K$,

$$\begin{aligned}
&\Pr\{\lim \kappa(\alpha) < k_0\} = 0, \\
&\Pr\{\lim \kappa(\alpha) = k_0\} \overset{asy.}{=} \Pr\{\max_{j>k_0} A_{k_0,j} \le \alpha\}, \\
&\Pr\{\lim \kappa(\alpha) > k_0\} \overset{asy.}{=} \Pr\{\max_{j>k_0} A_{k_0,j} > \alpha\}.
\end{aligned}$$

2) If $\alpha \to \infty$ and $r < \tau$,

$$\begin{aligned}
&\Pr\{\lim \kappa(\alpha) = k_0\} = 1, \\
&\Pr\{\lim \kappa(\alpha) \neq k_0\} = 0.
\end{aligned}$$

3) If $r > \tau$,

$$\begin{aligned}
&\Pr\{\lim \kappa(\alpha) < k_0\} = 1, \\
&\Pr\{\lim \kappa(\alpha) \ge k_0\} = 0.
\end{aligned}$$

4) If $r = \tau$,

$$\Pr\{\lim \kappa(\alpha) \le k_0\} = 1.$$

If $\mathcal{S}_k$ are nested, from (2.9),

$$\begin{bmatrix} A_{k_0,k_0+1} \\ A_{k_0,k_0+2} \\ \vdots \\ A_{k_0,K} \end{bmatrix} = \begin{bmatrix} 1 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ \vdots & \vdots & \ddots & \\ \frac{1}{K-k_0} & \frac{1}{K-k_0} & \cdots & \frac{1}{K-k_0} \end{bmatrix} \begin{bmatrix} Z_{k_0}/Z_K \\ Z_{k_0+1}/Z_K \\ \vdots \\ Z_{K-1}/Z_K \end{bmatrix} \triangleq \mathbf{G}_2 \mathbf{Z}_{k_0}/Z_K,$$

where $Z_j \sim \chi^2_1$, $j = k_0, \ldots, K-1$, and Z_K are independent. Define an indicator function $I(\mathbf{Z}_{k_0}, Z_K, \alpha) = I\{\max_{j>k_0} A_{k_0,j} \leq \alpha\}$. Then for $k_0 < K$, from Theorem 2.1,

$$\Pr\{\lim \kappa(\alpha) = k_0\} \overset{asy.}{=} E\{I(\mathbf{Z}_{k_0}, Z_K, \alpha)\} \overset{asy.}{=} E\{I(\mathbf{Z}_{k_0}, 1 + \delta^2/\sigma^2, \alpha)\}.$$

If there are no misspecified predictors, $\delta^2 = 0$.

If $\mathcal{S}_k$ are unnested, similar to (2.14), we may show asymptotically

$$\Pr\{\lim \kappa(\alpha) = k_0\} \leq E\{I(\mathbf{Z}_{k_0}, 1 + \delta^2/\sigma^2, \alpha)\}.$$

Hence for a finite or fixed α, $\Pr\{\lim \kappa(\alpha) = k_0\} \leq \Pr\{Z_{k_0} \leq (1 + \delta^2/\sigma^2)\alpha\} < 1$. This means that the FPE_α with a finite α is inconsistent if $k_0 < K$.

If $k_0 = K$, since we define $\max_{j>K} Z_{K,j} = 0$,

$$\Pr\{\lim \kappa(\alpha) = k_0\} = \Pr\{\max_{j>K} Z_{K,j} \leq \alpha\} = 1.$$

So in this extreme case, the FPE_α with a finite α is consistent.

The asymptotic properties are described in Corollaries 2.1 and 2.2, which are directly derived from Theorem 2.1. In the following corollaries, Assumptions 2.1 and 2.2 are implied but omitted.

Corollary 2.1. *Assume $k_0 < K$ and $\mathcal{S}_{k_0} \in \{\mathcal{S}_k\}$. Let $\alpha = \alpha_n$ and $n^{-1}\alpha_n \to r \neq \tau$. 1) If α is bounded, FPE_α is inconsistent, and the selected model is asymptotically either optimal or overfitted. 2) FPE_α is strongly consistent if and only if α is unbounded and $r < \tau$. 3) FPE_α is asymptotically underfitted if and only if α is unbounded and $r > \tau$.*

Corollary 2.2. *Assume $k_0 = K$ and $\mathcal{S}_{k_0} \in \{\mathcal{S}_k\}$. Let $n^{-1}\alpha_n \to r \neq \tau$. Then if $r < \tau$, FPE_α is strongly consistent. Otherwise if $r > \tau$, FPE_α is asymptotically underfitted.*

From Corollary 2.1, as commonly known, FPE_α is inconsistent if α is bounded. But the usual condition $n^{-1}\alpha_n \to 0$ is sufficient but not necessary for the consistency. From Corollary 2.2, AIC and FPE may be consistent in the extreme case of $k_0 = K$. So strictly speaking, the statement that AIC and FPE are inconsistent is incorrect.

The τ is an unknown constant related to the optimal model size k_0, and might be estimated by $\hat{\tau} = \min_{j<k_0} A_{k_0}(j)/n$ if k_0 was known. Since $A_{k_0}(j)/n \xrightarrow{a.s.} 0$ for $j > k_0$, the τ may be estimated by

$$\hat{\tau} = \min_{j<k}\{A_k(j)/n > \epsilon\}, \tag{2.18}$$

where ϵ is a small value.

2.5 Adaptive Procedures Based on Probabilities

2.5.1 Procedure One

According to the probability distribution of selecting a model, we may decide the best model. From the asymptotic property given in Theorem 2.1, we see that as the tuning parameter α increases with the sample size, the asymptotic probability of selecting the optimal model, $\mathcal{S}_{k_0}$, may approach one. Hence the model having the maximum probability should be the best model.

Let $p_k(\alpha) = \Pr\{\kappa(\alpha) = k\}$, the probability of selecting a model that can be estimated using (2.8), (2.10), or (2.13). The best model is estimated by

$$\hat{k}_0 = \arg \max_{1\le k\le K}\{\max_{\alpha_1\le\alpha\le\alpha_2} p_k(\alpha)\}, \tag{2.19}$$

where $[\alpha_1,\ \alpha_2]$ is a specified range of α. From the bound probability of selecting the optimal model (2.16) and Figure 2.1, the α may be limited by $2 \le \alpha \le 9$.

The model selection procedure given by (2.19) is called an adaptive FPE_α , denoted as $\text{FPE}_{\alpha,d}$, in which the adaptive change of α is based on the probability distribution. From Theorem 2.1, the $\text{FPE}_{\alpha,d}$ is inconsistent if the upper range α_2 is finite, and it

is consistent if $\alpha_2 = c\log(n)$, where c is a constant. In practice, the α may only take the integers in $[\alpha_1,\ \alpha_2]$ because whether or not a model can be selected relies on an interval in which α is.

2.5.2 Procedure Two

Let $I_k = [\max_{j>k} A_{k,j},\ \min_{j<k} A_{k,j}]$. If $\alpha \in I_k$, FPE_α selects the model $\mathcal{S}_k$. Let $\alpha = \alpha_{i_k} \in I_k$ be changed adaptively based on the interval. From Lemma 2.1, the probability of selecting the model $\mathcal{S}_k$ is

$$\begin{aligned}\Pr\{\kappa(\alpha_{i_k}) = k\} &= \Pr\{\max_{j>k} A_{k,j} \le \min_{j<k} A_{k,j}\} \\ &= \Pr\{\max_{j>k} D_{k,j} \le \min_{j<k} D_{k,j}\} \\ &= \Pr\{\max_{j>k} W_{k,j} \le \min_{j<k} W_{k,j}\},\end{aligned}$$

where $D_{k,j} = (\text{RSS}_k - \text{RSS}_j)/(j-k)$ and $W_{k,j} = D_{k,j}/\sigma^2$.

Similar to (2.14), for $k_0 < k < K$,

$$\begin{aligned}\Pr\{\kappa(\alpha_{i_k}) = k\} &\le \Pr\{W_{k,k+1} \le W_{k,k-1}\} \\ &\le \Pr\{\chi^2_1 \le \chi^2_1\} \\ &= 0.50.\end{aligned}$$

Under Assumptions 2.1 and 2.2, from Lemma 2.3, for $j > k_0 > l$, as n is large enough, we have $D_{k_0,j} = o(n) < D_{k_0,l} = O(n)$, and then

$$\Pr\{\kappa(\alpha_{i_{k_0}}) = k_0\} = \Pr\{\max_{j>k_0} D_{k_0,j} \le \min_{j<k_0} D_{k_0,j}\} = 1.$$

Hence by computing the probabilities, $p_k = \Pr\{\kappa(\alpha_{i_k}) = k\}$, $k = 1, \ldots, K-1$, we may determine the best model. The procedure is given by

$$\hat{k}_0 = \max\{k : p_k > \eta\}, \tag{2.20}$$

where η is the level of selecting the optimal model. If it is desired that $p_{k_0} \ge 0.90$, then $\eta = 0.90$. The procedure given in (2.20), denoted as $\text{FPE}_{\alpha,i}$, is consistent if $\eta \ge 0.5$.

Define an indicator function $I(X, y, k) = I\{\max_{j>k} D_{k,j} \leq \min_{j<k} D_{k,j}\}$. Then

$$\Pr\{\kappa(\alpha_{i_k}) = k\} = E\{I(X, y, k)\}.$$

Let $X^{(i)}$ and $y^{(i)}$ be the sample of X and y. By strong law of large numbers,

$$p_k = \frac{1}{N}\sum_{i=1}^{N} I(X^{(i)}, y^{(i)}, k) \xrightarrow{a.s.} \Pr\{\kappa(\alpha_{i_k}) = k\}. \tag{2.21}$$

The bootstrap or resampling methods can be used to estimate p_k.

2.6 Further Discussions

A more general form of FPE_α is

$$\text{FPE}_\alpha = \text{RSS}_k + \alpha c_k s_K^2, \tag{2.22}$$

where c_k is a known positive increasing function of the model size k and may depend on n. The c_k represents the model complexity. The form (2.22) includes a variety of information criteria. With $c_k = (e^{\alpha k/n} - 1)\text{RSS}_k/\alpha s_K^2$, the FPE_α is exactly equivalent to AIC_α. If $\alpha c_k = 4\sum_{j=1}^{k} \log(K/j)$, the FPE_α is the covariance inflation criterion (CIC) (Tibshirani and Knight, 1999). The FPE_α also includes the risk inflation criterion (RIC) (Foster and George, 1994).

In (2.22), s_K^2, an estimate of variance σ^2, may be considered as a scale of the tuning parameter α and can be replaced by any positive number. Substituting αs_K^2 with γ, we obtain,

$$\text{FPE}_\gamma = \text{RSS}_k + \gamma c_k. \tag{2.23}$$

The penalty in (2.22) is a random variable related to the sample, while the penalty in (2.23) is deterministic. The FPE_α is asymptotically equivalent to FPE_γ with $\gamma = (\sigma^2 + \delta^2)\alpha$. In the high dimensional feature space, usually s_K^2 is close or equal to zero, and then FPE_α cannot work but FPE_γ works.

Corollary 2.3. FPE_α *, defined in (2.22), selects the model* $\mathcal{S}_k$ *if and only if*

$$\max_{j>k} B_{k,j} \leq \alpha \leq \min_{j<k} B_{k,j},$$

where $B_{k,j} = (\text{RSS}_k - \text{RSS}_j)/\{(c_j - c_k)s_K^2\}$ *for* $j \neq k$, $k = 1, \ldots, K$. *Let* k_i, $i = 1, ..., m$, *be all of sizes selected by the* FPE_α *with different* $\alpha = \alpha_i$ *and be in ascending order. Then* $B_{k_1,k_2} \leq \alpha_1 < \infty$,

$$B_{k_i,k_{i+1}} \leq \alpha_i \leq B_{k_{i-1},k_i},$$

and $0 \leq \alpha_m \leq B_{k_{m-1},k_m}$. *If* $\alpha = B_{k_i,k_{i+1}}$, FPE_α *may select* $\mathcal{S}_{k_i}$ *and* $\mathcal{S}_{k_{i+1}}$.

Corollary 2.4. FPE_γ *, defined in (2.23), selects the model* $\mathcal{S}_k$ *if and only if*

$$\max_{j>k} D_{k,j} \leq \gamma \leq \min_{j<k} D_{k,j},$$

where $D_{k,j} = (\text{RSS}_k - \text{RSS}_j)/(c_j - c_k)$ *for* $j \neq k$, $k = 1, \ldots, K$. *Let* k_i, $i = 1, ..., m$, *be all of sizes selected by* FPE_γ *with different* $\gamma = \gamma_i$ *and in ascending order. Then* $D_{k_1,k_2} \leq \gamma_1 < \infty$,

$$D_{k_i,k_{i+1}} \leq \gamma_i \leq D_{k_{i-1},k_i},$$

and $0 \leq \gamma_m \leq D_{k_{m-1},k_m}$.

Corollaries 2.3 and 2.4 are directly obtained from Lemma 2.1. Substituting $A_{k,j}$ with $B_{k,j}$ or $D_{k,j}$, the results provided in Section 2.3 and Section 2.4 are applicable to the general FPE_α or FPE_γ.

From Corollaries 2.3 and 2.4, the necessary condition under which the model $\mathcal{S}_k$ can be selected is

$$D_{k_i,k_{i+1}} \leq D_{k_{i-1},k_i}.$$

Let $k_0 < k_1 < \ldots < k_m$. If $\mathcal{S}_{k_0} \subset \mathcal{S}_{k_1} \subset \ldots \subset \mathcal{S}_{k_m}$ and let $c_k = k$ then $D_{k_i,k_{i+1}}/\sigma^2 \sim \chi^2_{k_{i+1}-k_i}/(k_{i+1} - k_i)$ are independent, see Lemma 2.2, and the probability that the m models could be selected is

$$\Pr\left\{\frac{\chi^2_{k_1-k_0}}{k_1 - k_0} \leq \ldots \leq \frac{\chi^2_{k_m-k_{m-1}}}{k_m - k_{m-1}}\right\},$$

which will be near zero if m is large. Hence only a few models can be selected even if K is large. Using the interval constraints may significantly reduce the number of the candidate models. This is particularly useful to reduce the computational burden if K is large.

2.7 Numerical Illustration

This section is to numerically illustrate the properties of the general FPE_α and FPE_γ with simulation study and applications to the diabetes study and the Standard and Poors 500 stocks. The simulation study verifies the probability of selecting a model by FPE_γ and the probability-based procedures for model selection. For the diabetes study, we obtain the set of candidate models by LASSO and exhaustive search, and then use the probability-based procedures to select the best model. In the dataset of the Standard and Poors 500 stocks, the number of observations is less than the number of the predictors. We obtain the candidate models by MCP , screen the candidate models by the interval conditions in Corollary 2.4, and then use the probability-based procedures to select the final best model. In this section, $c_k = k$.

2.7.1 Simulation Study

As in Tibshirani (1996) and Fan and Li (2001), we consider a linear model with $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$. The components of X and ε are standard normal. The correlation between X_i and X_j is $\rho^{|i-j|}$ with $\rho = 0.5$. The true model size is $k_0 = 3$. The full model size is $K = 8$. Let $\sigma = 1$, and $n = 20$.

First we assume that the predictors are deterministic, that is, X_k, $k = 1, \ldots, 8$, are fixed. We run $m = 10^4$ simulations, and count the number of the models selected by FPE_γ with different $\gamma = \sigma^2\alpha_\sigma$, $\alpha_\sigma = 2, ..., 10$. Let m_k be the number of the selected model $\mathcal{S}_k$, $k = 1, ..., 8$. The relative frequency, defined by $f_k = m_k/m$, the model error, $\text{ME}(\hat{\beta}) = \| X\beta - X(\kappa(\alpha))\hat{\beta}(\kappa(\alpha)) \|^2$, and the standard error (SE) of the ME are shown in Table 2.1. FPE_γ with $\alpha_\sigma \in (7, 10)$ selects the true model with probability more than 0.99. FPE_γ with $\alpha_\sigma = 9$ has the highest probability but slightly larger ME than FPE_γ with $\alpha_\sigma = 8$. This means that the true model having the minimum model error may not be the model having the minimum prediction error.

With known noncentrality parameters $\lambda_{i,i+1}$, we compute the probabilities p_k using (2.8) with $N = 10^7$. The results are in Table 2.2. The largest standard error is 0.00014. Comparing with Table 2.1, the probabilities and the relative frequencies are very close, and their differences are from -0.0062 to 0.0053.

Now we assume that the predictors are random variables. we compare the performance of the probability-based procedures $\text{FPE}_{\alpha,d}$ and $\text{FPE}_{\alpha,i}$ with AIC and BIC by measuring the percentage number of underfit, overfit and correct models, and the

Table 2.1: Relative frequency f_k of selecting a model $\mathcal{S}_k$, $k = 1, ..., 8$, by FPE$_\gamma$.

α_σ	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	ME	SE
2	0	0.000	0.741	0.112	0.058	0.039	0.026	0.024	4.722	0.043
3	0	0.000	0.884	0.073	0.025	0.012	0.004	0.002	3.802	0.035
4	0	0.000	0.946	0.042	0.008	0.003	0.000	0.000	3.410	0.031
5	0	0.000	0.970	0.025	0.004	0.001	0.000	0.000	3.259	0.030
6	0	0.001	0.982	0.016	0.001	0.000	0.000	0.000	3.170	0.028
7	0	0.001	0.990	0.008	0.000	0.000	0.000	0.000	3.123	0.028
8	0	0.002	0.993	0.004	0.000	0.000	0.000	0.000	**3.109**	0.029
9	0	0.003	**0.994**	0.003	0.000	0.000	0.000	0.000	3.121	0.031
10	0	0.006	0.993	0.002	0.000	0.000	0.000	0.000	3.172	0.034

Table 2.2: The probability of selecting a model $\mathcal{S}_k$, $k = 1, ..., 8$, with known noncentrality parameters. The standard errors are between 0 and 0.00014. The differences, $p_k - f_k$, range from -0.0062 to 0.0053.

α_σ	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
2	0	0.000	0.735	0.117	0.061	0.038	0.027	0.022
3	0	0.000	0.883	0.074	0.025	0.011	0.005	0.003
4	0	0.000	0.943	0.043	0.010	0.003	0.001	0.000
5	0	0.000	0.971	0.025	0.004	0.001	0.000	0.000
6	0	0.001	0.984	0.014	0.001	0.000	0.000	0.000
7	0	0.001	0.990	0.008	0.000	0.000	0.000	0.000
8	0	0.002	**0.993**	0.005	0.000	0.000	0.000	0.000
9	0	0.004	0.993	0.003	0.000	0.000	0.000	0.000
10	0	0.006	0.992	0.002	0.000	0.000	0.000	0.000

model error, $(\hat{\beta} - \beta)'\Sigma_x(\hat{\beta} - \beta)$, where Σ_x is the covariance matrix of the predictors. The probabilities $p_k(\alpha)$ and p_k are computed using (2.8) and (2.21) by bootstrap methods. The level of selecting the optimal model for FPE$_{\alpha,i}$ is $\eta = 0.90$. The number of bootstrap samples is $N = 100$. We run 10^4 simulations. The results are in Table 2.3.

2.7.2 Diabetes Study

We consider the diabetes dataset used by Efron et al. (2004) to illustrate LARS. This dataset consists of 442 diabetes patients who were measured on 10 baseline variables: age, sex, body mass index, average blood pressure and six blood serum measurements. A prediction model was desired for the response variable, a quantitative measure of disease progression one year after baseline.

Table 2.3: Percentage number of underfitted models (u), correct models (c), and overfitted models (o), and model error with standard deviation from 10^4 simulations for different sample size n.

n	procedure	u	c	o	me	sd
50	AIC	0	44	56	0.304	0.245
	BIC	0	78	22	0.220	0.215
	$\text{FPE}_{\alpha,d}$	0	94	6	0.179	0.207
	$\text{FPE}_{\alpha,i}$	4	95	1	0.217	0.413
100	AIC	0	44	56	0.140	0.105
	BIC	0	86	14	0.094	0.086
	$\text{FPE}_{\alpha,d}$	0	97	3	0.078	0.073
	$\text{FPE}_{\alpha,i}$	0	99	1	0.073	0.065

We first consider the subsets generated by LASSO , which consist of ten nested subsets $\mathcal{S}_k$, $k = 1, \ldots, 10$. The best model will be selected from the subsets. We compute the probability of selecting each model by (2.10). Monte Carlo sample size $N = 10^6$. The results are shown in Table 2.4. The model $\mathcal{S}_5$ is selected using the procedure (2.19), $\text{FPE}_{\alpha,d}$, because it has the highest probability. Using the C_p in LARS selects the model $\mathcal{S}_7$ that contains two insignificant coefficients. AIC also selects $\mathcal{S}_7$ but BIC prefers $\mathcal{S}_5$.

Table 2.4: The probability of selecting a model $\mathcal{S}_k$, $k = 1, ..., 10$, in the LASSO subsets, computed by (2.10). $N = 1000,000$. The standard errors are between 0 and 0.0005.

α	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
2	0	0.000	0.000	0.003	0.333	0.038	0.353	0.069	0.147	0.057
3	0	0.000	0.003	0.012	0.550	0.032	0.296	0.033	0.061	0.013
4	0	0.000	0.010	0.025	0.695	0.023	0.207	0.014	0.022	0.003
5	0	0.001	0.025	0.043	0.772	0.015	0.131	0.005	0.007	0.001
6	0	0.004	0.049	0.062	**0.794**	0.009	0.077	0.002	0.002	0.000
7	0	0.010	0.084	0.080	0.776	0.005	0.043	0.001	0.001	0.000
8	0	0.020	0.128	0.095	0.730	0.003	0.023	0.000	0.000	0.000
9	0	0.037	0.177	0.107	0.664	0.002	0.012	0.000	0.000	0.000
10	0	0.063	0.228	0.113	0.587	0.001	0.006	0.000	0.000	0.000

We further consider the best subsets generated by exhaustive search. The best subsets are not nested, see Table 2.7, where five subsets are listed. The $\mathcal{S}_5$ is the same as that in LASSO subsets, but the $\mathcal{S}_6$ is different from that in LASSO. In the best subsets, $\mathcal{S}_5$ and $\mathcal{S}_6$ are almost equivalent because the covariate 'hdl' in $\mathcal{S}_5$ has a

linear relationship with covariates 'tc' and 'ldl' in $\mathcal{S}_6$. The probabilities of selecting the models $\mathcal{S}_k$ are computed using (2.13) and shown in Table 2.5. It is seen that $\mathcal{S}_5$ and $\mathcal{S}_6$ are tied and both can be selected using the procedure (2.19). Using Lemma 2.1, we calculate the intervals $(\alpha_{k,1}, \alpha_{k,2})$ shown in Table 2.6. From Table 2.6 BIC selects the model $\mathcal{S}_5$ because $\log n = 6.09$, and AIC selects the model $\mathcal{S}_6$.

Table 2.5: The probability of selecting a model $\mathcal{S}_k$, $k = 1, ..., 10$, in the best subsets, computed by (2.13). $N = 1000,000$. The standard errors are between 0 and 0.0005.

α	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
2	0	0.000	0.000	0.002	0.094	0.386	0.182	0.193	0.089	0.053
3	0	0.000	0.001	0.008	0.182	0.503	0.140	0.116	0.036	0.013
4	0	0.000	0.002	0.023	0.277	**0.529**	0.091	0.061	0.014	0.003
5	0	0.000	0.007	0.045	0.367	0.492	0.053	0.029	0.005	0.001
6	0	0.001	0.016	0.074	0.442	0.423	0.028	0.013	0.002	0.000
7	0	0.003	0.032	0.105	0.495	0.343	0.014	0.005	0.001	0.000
8	0	0.008	0.056	0.136	0.527	0.264	0.007	0.002	0.000	0.000
9	0	0.018	0.088	0.160	**0.535**	0.195	0.003	0.001	0.000	0.000
10	0	0.032	0.126	0.178	0.523	0.138	0.001	0.000	0.000	0.000

Table 2.6: Intervals $(\alpha_{k,1}, \alpha_{k,2})$ in which FPE_α selects the model $\mathcal{S}_k$.

$\mathcal{S}_k$	1	2	3	**5**	**6**	7	8	9	10
$\alpha_{k,1}$	103.52	18.45	12.79	**5.60**	**1.26**	1.06	0.22	0.03	0
$\alpha_{k,2}$	∞	103.52	18.45	**12.79**	**5.60**	1.26	1.06	0.22	0.03

Table 2.7: Model coefficients with significance codes '***', '**', '*', '.', or '-' representing the corresponding p-value in $(0, 0.001]$, $(0.001, 0.01]$, $(0.01, 0.05]$, $(0.05, 0.1]$, or $(0.1, 1]$, respectively.

Variable	$\mathcal{S}_3$	$\mathcal{S}_5$	$\mathcal{S}_6$	$\mathcal{S}_7$	$\mathcal{S}_8$
sex		-0.1456***	-0.1399***	-0.1462***	-0.1495***
bmi	0.3725***	0.3234***	0.3273***	0.3268***	0.3204***
map	0.1620***	0.2015***	0.2021***	0.2062***	0.1986***
tc			-0.4682***	-0.3799**	-0.3834**
ldl			0.3327***	0.2181 -	0.2186 -
hdl		-0.1786***			
tch				0.0836 -	0.0786 -
ltg	0.3359***	0.2930***	0.4967***	0.4357***	0.4274***
glu					0.0415 -

2.7.3 Standard & Poor's 500 Index

The dataset included in R package *plus* contains a year's worth of close-of-day data for most of the Standard and Poors 500 stocks. We consider the daily percentage change. In the data, the first column named X.DJI, Dow Jones Industrial Average, is used for the response variable, and the other 492 columns excluding the second column are the predictor variables. There are 252 observations. The goal is to estimate the index using the individual stocks.

In this example, $n = 252$ and $d_n = 492$. First, the minimax concave penalized likelihood method, MCP (Zhang, 2010), is used to obtain 501 candidate models with sizes from 1 to 30. Then, we select a best model having a minimum RSS for each size and get $K = 28$ candidate models with unique size from 1 to 30. Using Corollary 2.4, we calculate the intervals of γ and then reduce the 28 candidate models to 10 that can be possibly selected by FPE_γ, see Table 2.8. Finally, using the probability-based procedure (2.19) yields the best model $\mathcal{S}_{18}$ with a probability of more than 0.95.

Table 2.8: Intervals $(\gamma_{k,1}, \gamma_{k,2})$ in which FPE_γ selects the model $\mathcal{S}_k$.

$\mathcal{S}_k$	1	3	4	7	11	13	**18**	22	24	30
$\gamma_{k,1}$	104.00	18.25	14.36	3.07	1.46	1.13	**0.28**	0.277	0.206	0
$\gamma_{k,2}$	∞	104.00	18.25	14.36	3.07	1.46	**1.13**	0.28	0.277	0.206

2.8 Conclusions

Model selection is to find the best model among a set of candidate models according to some selection criterion. We considered the generalized information criterion, FPE_α, investigated the relevant properties, and extended it into FPE_γ that is suitable for high dimensional space. We investigated the probability distribution of selecting a model by FPE_α, and the conditions under which the criterion is underfitting, consistent, or overfitting. The probability-based procedures were proposed for selecting the best model.

As shown in the application to the Standard and Poors 500 stocks, the properties given in this chapter have a potential application to the model selection in high dimensional feature space. First, using the necessary and sufficient conditions that the models can be selected reduces the number of the candidate models. Then, using

the probability-based procedures select the best model in the reduced set of candidate models.

2.9 Appendix

2.9.1 Proofs of Lemmas

Proof of Lemma 2.1.

FPE$_\alpha$ selects the model $\mathcal{S}_k$ if and only if for $j \neq k$, FPE$_\alpha(j)$ − FPE$_\alpha$ = (RSS$_j$ − RSS$_k$) + $\alpha(j-k)s_K^2 \geq 0$, that is, $\max_{j>k} A_{k,j} \leq \alpha \leq \min_{j<k} A_{k,j}$.

Let $a_{1k} = \max_{j>k} A_{k,j}$ and $a_{2k} = \min_{j<k} A_{k,j}$. Since

$$a_{1k_i} = \max_{j>k_i} F_{k_i}(j) \geq A_{k_i,k_{i+1}} = F_{k_{i+1}}(k_i) \geq \min_{j<k_{i+1}} F_{k_{i+1}}(j) = a_{2k_{i+1}},$$

and there is no gap between the adjacent intervals $[a_{1k_i}, a_{2k_i}]$ and $[a_{1k_{i+1}}, a_{2k_{i+1}}]$, we have $a_{1k_i} = a_{2k_{i+1}}$, and then $a_{1k_i} = A_{k_i,k_{i+1}} = F_{k_{i+1}}(k_i) = a_{2k_{i+1}}$.

If $\alpha = A_{k_i,k_{i+1}}$, $\alpha \in [a_{1k_i}, a_{2k_i}]$ and $\alpha \in [a_{1k_{i+1}}, a_{2k_{i+1}}]$. Hence FPE$_\alpha$ can select the two models $\mathcal{S}_{k_i}$ and $\mathcal{S}_{k_{i+1}}$. □

Lemma 2.4. *Under Assumption 2.2, there exist the following limits: (1)* $n^{-1}X_i'\varepsilon \xrightarrow{a.s.} 0$, $n^{-1}\mu'\varepsilon \xrightarrow{a.s.} 0$, *and* $n^{-1}X(k)'\mu \xrightarrow{a.s.} v_k$; *(2)* $n^{-1}\mu_a'\mu_a \xrightarrow{a.s.} v_{k_0}'V_{k_0k_0}^{-1}v_{k_0}$, *and* $n^{-1}\mu_e'\mu_e \xrightarrow{a.s.} \delta^2$, *where* $\delta^2 = v^2 - \sigma^2 - v_{k_0}'V_{k_0k_0}^{-1}v_{k_0}$.

Proof of Lemma 2.4.

(1) Since $Z_n = n^{-1}X_i'\varepsilon \sim N(0, \sigma^2 X_i'X_i/n^2)$, and there exist N_0 and σ_i such that $X_i'X_i/n < {\sigma_i}^2$ for $n > N_0$, for any $\epsilon > 0$

$$\begin{aligned} Pr\{|Z_n| > \epsilon\} &= \textstyle\int_{|z|>n\epsilon/\sigma\|X_i\|} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &\leq \textstyle\int_{|z|>\sqrt{n}\epsilon/\sigma\sigma_i} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &\leq \textstyle\int_{|z|>\sqrt{n}\epsilon/\sigma\sigma_i} \frac{1}{\sqrt{2\pi}} \frac{2^k k!}{z^{2k}} dz = \frac{c_k}{n^{k-0.5}}, \ (k \geq 2). \end{aligned}$$

Since $\sum_{n>N_0} Pr\{|Z_n| > \epsilon\} < \infty$, $Z_n \xrightarrow{a.s.} 0$. Similarly, $n^{-1}\mu'\varepsilon \xrightarrow{a.s.} 0$. Then $\lim n^{-1}X(k)'\mu = \lim n^{-1}X(k)'(y-\varepsilon) = \lim n^{-1}X(k)'y = v_k, a.s.$

(2)

$$\begin{aligned}\lim n^{-1}\mu_a'\mu_a &= \lim n^{-1}\mu' H_{k_0}\mu \\ &= \lim\{n^{-1}\mu' X(k_0)\}\{n^{-1}X(k_0)'X(k_0)\}^{-1}\{n^{-1}X(k_0)'\mu\} \\ &= v_{k_0}' V_{k_0k_0}^{-1} v_{k_0}.\end{aligned}$$

Since $\lim n^{-1}y'y = \lim n^{-1}\mu'\mu + \lim n^{-1}\varepsilon'\varepsilon$, $\lim n^{-1}\mu'\mu = v^2 - \sigma^2$, and then $\lim n^{-1}\mu_e'\mu_e = \lim n^{-1}\mu'\mu - \lim n^{-1}\mu_a'\mu_a = \delta^2$. □

Proof of Lemma 2.3.

By $H_k\mu = H_k\mu_a$, and Lemma 2.4,

$$\begin{aligned}\Delta_k^2 &= \lim n^{-1}\| X(k_0)\beta(k_0) - X(k)\beta(k) \|^2 \\ &= \lim n^{-1}\| \mu_a - H_k\mu \|^2 \\ &= \lim(n^{-1}\mu_a'\mu_a - n^{-1}\mu' H_k\mu) \\ &= v_{k_0}' V_{k_0k_0}^{-1} v_{k_0} - v_k' V_{kk}^{-1} v_k.\end{aligned}$$

From

$$\begin{aligned}n^{-1}\text{ME}(\beta(k)) &= n^{-1}\| \mu - X(k)\beta(k) \|^2 \\ &= n^{-1}(\| \mu_a - X(k)\beta(k) \|^2 + \mu_e'\mu_e) \\ &= n^{-1}(\| X(k_0)\beta(k_0) - X(k)\beta(k) \|^2 + \mu_e'\mu_e),\end{aligned}$$

and Lemma 2.4, $n^{-1}\text{ME}(\beta(k)) \xrightarrow{a.s.} \Delta_k^2 + \delta^2$.

The second limit follows from Lemma 2.4 and

$$\text{RSS}(\beta(k)) = \text{ME}(\beta(k)) + 2\mu'(I - H_k)\varepsilon + \varepsilon'(I - H_k)\varepsilon.$$

If $k < k_0$, by Assumption 2.1, $\Delta_k^2 > 0$. If $k \geq k_0$, $k \geq k_0$ as n is large enough. There exists $\tilde{\mathcal{S}}_k \supseteq \mathcal{S}_{k_0}$ and then $\Delta_k^2(\tilde{\mathcal{S}}_k) = 0$, where $k(\tilde{\mathcal{S}}_k) = k$. Since $\mathcal{S}_k$ is the best one in the models of size k, $\text{RSS}(\beta(k)) \leq \text{RSS}(\tilde{\mathcal{S}}_k)$. From Lemma 2.3,

$$\lim_n n^{-1}\text{RSS}(\beta(k)) = \Delta_k^2 + \delta^2 + \sigma^2 \leq \lim_n n^{-1}\text{RSS}(\tilde{\mathcal{S}}_k) = \delta^2 + \sigma^2.$$

Hence $\Delta_k^2 = 0$. □

2.9.2 Proofs of Propositions

Proof of Proposition 2.1.

Let $\text{SNR}(\mathcal{S}_k) = \| X(k_0)\beta(k_0) - X(k)\beta(k) \|^2 / \sigma^2$. Then

$$\begin{aligned} \text{PE}(\mathcal{S}_k) &= \text{ME}(\beta(k)) + k\sigma^2 + n\sigma^2 \\ &= \| \mu_a - X(k)\beta(k) \|^2 + \mu_e'\mu_e + k\sigma^2 + n\sigma^2 \\ &= \| X(k_0)\beta(k_0) - X(k)\beta(k) \|^2 + k\sigma^2 + \mu_e'\mu_e + n\sigma^2 \\ &= \text{PE}(\mathcal{S}_{k_0}) + \{\text{SNR}(\mathcal{S}_k) - (k_0 - k)\}\sigma^2. \end{aligned} \tag{2.24}$$

For $k > k_0$, $\text{PE}(\mathcal{S}_k) > \text{PE}(\mathcal{S}_{k_0})$. So $k_{\text{PE}} \leq k_0$.

If $\text{SNR} > 1$, $\text{PE}(\mathcal{S}_k) > \text{PE}(\mathcal{S}_{k_0})$ for $k \neq k_0$. Hence $k_{\text{PE}} = k_0$ and $\text{PE}(\mathcal{S}_{k_{\text{PE}}}) = \text{PE}(\mathcal{S}_{k_0})$. From (2.24), $\text{SNR}(\mathcal{S}_{k_{\text{PE}}}) = 0$ and then

$$X(k_{\text{PE}})\beta(k_{\text{PE}}) = X(k_0)\beta(k_0) = \mu_a.$$

By Assumption , $\mathcal{S}_{k_{\text{PE}}} = \mathcal{S}_{k_0}$.

If $\text{SNR} < 1$, From (2.24), there exists $k_1 < k_0$ such that $\text{PE}(\mathcal{S}_{k_1}) < \text{PE}(\mathcal{S}_{k_0})$. Hence $k_{\text{PE}} \neq k_0$, and then $k_{\text{PE}} < k_0$. This completes the proof. □

Proof of Proposition 2.3.

Consider the models $\mathcal{S}$ with $\kappa(\mathcal{S}) = k$. From Lemma 2.3, as $n \to \infty$, the $\mathcal{S}_k$ minimizing $\text{RSS}(\mathcal{S})$ almost surely approaches the model minimizing Δ_k that is independent of n.

Similarly, as $n \to \infty$, the $\mathcal{S}_k$ minimizing $\text{RSS}(\mathcal{S})$ subject to $\kappa(\mathcal{S}) = k_0$ almost surely approaches the model minimizing Δ_k that is $\mathcal{S}_{k_0}$. Hence $\mathcal{S}_{k_0} \in \{\mathcal{S}_k\}$.

For the fixed $k < j$, $\mathcal{S}_k$ and $\mathcal{S}_j$ are almost surely independent of the sample as n is large enough. Since $\text{RSS}_k \geq \text{RSS}_j$, $\text{RSS}_k - \text{RSS}_j = y'(H_j - H_k)y \geq 0$ for each sample y. So either $H_j y = H_k y$ or $H_j - H_k \geq \mathbf{0}$ and then $\mathcal{S}_k \subset \mathcal{S}_j$. □

2.9.3 Proofs of Theorems

Proof of Theorem 2.1.

Let n be large enough. Then from Proposition 2.3, $\mathcal{S}_{k_0} \in \{\mathcal{S}_k\}$. For $j, k \geq k_0 > l$, from Lemma 2.3, $A_{k,j} = o(n)$ and $A_{k,l} = O(n)$. Hence $A_{k_0,j} < A_{k_0,l}$ and

$$\alpha_{k_0,1} = \max_{j > k_0} A_{k_0,j} < \alpha_{k_0,2} = \min_{l < k_0} A_{k_0,l}.$$

So from Lemma 2.1, the optimal model $\mathcal{S}_{k_0}$ can be always selected by the FPE_α with a proper α. We may divide $[0, \infty)$ into three intervals:

$$\begin{aligned}\mathcal{A}_- &= [\alpha_{k_0,2},\ \infty),\\ \mathcal{A}_0 &= [\alpha_{k_0,1},\ \alpha_{k_0,2}),\\ \mathcal{A}_+ &= [0,\ \alpha_{k_0,1}).\end{aligned}$$

From Lemma 2.1,

$$\begin{aligned}\Pr\{\kappa(\alpha) < k_0\} &= \Pr\{\alpha \in \mathcal{A}_-\},\\ \Pr\{\kappa(\alpha) = k_0\} &= \Pr\{\alpha \in \mathcal{A}_0\},\\ \Pr\{\kappa(\alpha) > k_0\} &= \Pr\{\alpha \in \mathcal{A}_+\}.\end{aligned}$$

1) Let $\alpha < \infty$. Since $A_{k_0,l} = O(n)$ for $l < k_0$, $\alpha_{k_0,2} \to \infty$. As n is large enough,

$$\begin{aligned}\Pr\{\kappa(\alpha) < k_0\} &= \Pr\{\alpha \in \mathcal{A}_-\} = 0,\\ \Pr\{\kappa(\alpha) = k_0\} &= \Pr\{\alpha \in \mathcal{A}_0\} = \Pr\{\alpha \in [\alpha_{k_0,1},\ \infty)\},\\ \Pr\{\kappa(\alpha) > k_0\} &= \Pr\{\alpha \in \mathcal{A}_+\} = \Pr\{\alpha \in [0,\ \alpha_{k_0,1})\}.\end{aligned}$$

2) Let $\alpha_n \to \infty$ and $r < \tau$. Since $\lim n^{-1}\alpha_n = r < \tau = \lim \min_{l<k_0} n^{-1} A_{k_0,l}$, there exists n_0 such that for $n > n_0$, $\alpha_n < \min_{l<k_0} A_{k_0,l} = \alpha_{k_0,2}$. Hence

$$\Pr\{\kappa(\alpha) = k_0\} = \Pr\{\alpha \in \mathcal{A}_0\} = \Pr\{\max_{j>k_0} A_{k_0,j} \le \alpha_n\}.$$

Since $\Pr\{\max_{j>k_0} A_{k_0,j} \le \alpha_n\} \overset{asy.}{=} \Pr\{\max_{j>k_0} W_{k_0,j} \le \alpha_n(1 + \delta^2/\sigma^2)\}$, where $W_{k_0,j} = (\text{RSS}_{k_0} - \text{RSS}_j)/\sigma^2(j - k_0)$. From Proposition 2.3 and Lemma 2.2, as n is large enough, for $j > k_0$, either $W_{k_0,j} = 0$ or $W_{k_0,j} \sim \chi^2_{j-k_0}/(j - k_0)$. Hence from $\alpha_n \to \infty$, we have

$$\Pr\{\max_{j>k_0} W_{k_0,j} \le \alpha_n(1 + \delta^2/\sigma^2)\} \to 1,$$

and then $\Pr\{\kappa(\alpha) = k_0\} = 1$.

3) Let $r > \tau$. There exists n_0 such that for $n > n_0$, $\alpha > \min_{l<k_0} A_{k_0,l} = \alpha_{k_0,2}$. Hence

$$\Pr\{\kappa(\alpha) < k_0\} = \Pr\{\alpha \in \mathcal{A}_-\} = \Pr\{\min_{l<k_0} A_{k_0,l} \le \alpha\} = 1.$$

4) Let $r = \tau$.

$$\Pr\{\kappa(\alpha) > k_0\} = \Pr\{\alpha \in \mathcal{A}_+\} = \Pr\{\alpha/n < \alpha_{k_0,1}/n\} = \Pr\{r <= 0\} = 0.$$

Thus $\Pr\{\lim \kappa(\alpha) \leq k_0\} = 1$. □

Chapter 3

GIC WITH OVERFITTING LEVEL

A new model selection method using the generalized information criterion (GIC) is developed based on controlling the upper bound of the probability of selecting an overfitting model. The main advantage of this method is that it has very good model selection capability as well as being easy to implement. As in the case of the BIC, the new procedure is consistent with a proper choice of the overfitting level. The improvement in model selection over the BIC is demonstrated in simulation studies. The application of the GIC is illustrated with logistic regression and subset autoregression.

3.1 Introduction

Model selection is an important topic in modern applied statistics (Hastie et al., 2009, §7). We suggest using the generalized information criterion (GIC) to automatically select the best model from a set of candidate models. Many model selection criteria have been derived based on a variety of principles such as minimizing final prediction error (Akaike, 1969, 1970), minimizing mean squared model error (Mallows, 1973), minimizing information loss (Akaike, 1974), and maximizing posterior probability (Schwarz, 1978). Most of them may be considered as a special case of the GIC (Bhansali and Downham, 1977; Akaike, 1979; Shibata, 1984; Nishii, 1984; Shao, 1997; Zhang, 2009).

The GIC with hypothesis testing for model selection was considered for linear regression by Shao and Rao (2000). A model selection procedure was proposed by controlling an upper bound of overfitting probability to a pre-assigned level (Shao and Rao, 2000). This approach provides a computationally more efficient approach than cross-validation. Previously, Shao (1993, 1997) showed that simple leave-one-out and k-fold cross validation did not provide consistent model selection but that a more computationally intensive approach using delete-d cross-validation was needed

for asymptotic consistency. This approach was shown to provide more accurate model selection in finite samples (Shao, 1993, 1997) in simulation experiments.

The hypothesis testing approach of Shao and Rao (2000) is not easily implemented since the overfitting probability bound cannot be computed. As well, the suggested approximation (Shao and Rao, 2000, eqn (2.6)) is very conservative because of replacing the true model with a model of size one. In this chapter, we propose an alternative procedure by controlling the upper bound probability of selecting an overfitting model. The probability of selecting an overfitting model is different from the overfitting probability considered by Shao and Rao (2000). Our approach is simpler as well as more general.

3.2 Generalized Information Criterion

Let $y = (y_1, ..., y_n)$ be a vector of responses and $X = (X_1, ..., X_d)$ be an $n \times d$ matrix of inputs. Let $\mathcal{S} = \{s_1, ..., s_k\}$ be a subset of $\{1, 2, ..., d\}$, which represents a class of models with size k. The model is specified by a distribution function $f_{\theta(\mathcal{S})}(y|X(\mathcal{S}))$, where $\theta(\mathcal{S})$ is a vector of parameters, and $X(\mathcal{S})$ denotes the matrix formed by selecting the columns corresponding to $\mathcal{S}$ from X. After the data is available, let $L(\theta(\mathcal{S})) = f_{\theta(\mathcal{S})}(y|X(\mathcal{S}))$ be the likelihood function and $\hat{\theta}(\mathcal{S})$ its maximum likelihood estimate.

We consider the problem of selecting the best model from a set of candidate models denoted by $\{\mathcal{S}_k, k = 1, \ldots, K\}$, where $\mathcal{S}_k$ is the model of size k that has the maximum likelihood in the class of models with size k. Assume that the sizes of the candidate models are unique and $L(\hat{\theta}(\mathcal{S}_k)) < L(\hat{\theta}(\mathcal{S}_{k+1}))$. If the model sizes are not unique, we keep the model having the maximum likelihood and remove others for each size. Hence, selecting a model is equivalent to selecting the model size.

The best model is selected by the minimum value of some selection criterion. A widely used criterion is the generalized information criterion (Akaike, 1979; Nishii, 1984),

$$\text{GIC}_\alpha = -2 \log L(\hat{\theta}(\mathcal{S}_k)) + \alpha k. \tag{3.1}$$

When α is constant, the GIC_α was called AIC_α (Akaike, 1979; Bhansali, 1986), in which the α was introduced to balance the effects of the bias and variance of the parameter estimate. The GIC_α is called AIC-type if α is bounded and BIC-type if $\alpha \to \infty$ as $n \to \infty$. The consistent property of BIC is shared by the BIC-type criterion satisfying $\alpha \to \infty$ and $\alpha/n \to 0$ as $n \to \infty$ (Shao, 1997; Yang, 2005).

3.3 Procedure by Controlling Overfitting

Before giving the model selection procedure that is based on controlling the probability of selecting an overfit model, we analyze the constraints on the α under which the GIC_α can select a specified model. The GIC_α selects model $\mathcal{S}_k$ if and only if for $j \neq k$,

$$-2\{\log L(\hat{\theta}(\mathcal{S}_j)) - \log L(\hat{\theta}(\mathcal{S}_k))\} + \alpha(j-k) \geq 0.$$

Equivalently, for $j > k$,

$$\alpha \geq 2\{\log L(\hat{\theta}(\mathcal{S}_j)) - \log L(\hat{\theta}(\mathcal{S}_k))\}/(j-k),$$

and for $j < k$,

$$\alpha \leq 2\{\log L(\hat{\theta}(\mathcal{S}_j)) - \log L(\hat{\theta}(\mathcal{S}_k))\}/(j-k).$$

Hence, the following proposition holds.

Proposition 3.1. GIC_α *selects the model* $\mathcal{S}_k$ *if and only if* $\alpha_{k,1} \leq \alpha \leq \alpha_{k,2}$, *where*

$$\alpha_{k,1} = \max_{j>k} 2\{\log L(\hat{\theta}(\mathcal{S}_j)) - \log L(\hat{\theta}(\mathcal{S}_k))\}/(j-k),$$

$$\alpha_{k,2} = \min_{j<k} 2\{\log L(\hat{\theta}(\mathcal{S}_j)) - \log L(\hat{\theta}(\mathcal{S}_k))\}/(j-k).$$

Here we define $\alpha_{K,1} = 0$ and $\alpha_{1,2} = \infty$. Proposition 3.1 gives all of the possible models that can be selected by the GIC_α. The inequality condition $\alpha_{k,1} \leq \alpha_{k,2}$ may not hold for some k and in this case, the model $\mathcal{S}_k$ cannot be selected.

3.3.1 Hypotheses

Assume that the true model $\mathcal{S}_{k_0}$ is in the candidate models and $\mathcal{S}_{k_0} \subset \mathcal{S}_l$ for $k_0 < l$. Let $\kappa(\text{GIC}_\alpha)$ be the model size selected by GIC_α and $\kappa(\text{GIC}_\alpha) < K$. We consider the following hypotheses for testing overfitting

$$\begin{aligned} H_0 &: \ \kappa(\text{GIC}_\alpha) > k_0, \\ H_1 &: \ \kappa(\text{GIC}_\alpha) \leq k_0. \end{aligned}$$

Let $k = \kappa(\text{GIC}_\alpha)$. Let $\widetilde{\mathcal{S}}_{k-1}$ and $\widetilde{\mathcal{S}}_{k+1}$ be the models of size $k-1$ and $k+1$, respectively, and satisfy $\widetilde{\mathcal{S}}_{k-1} \subset \mathcal{S}_k \subset \widetilde{\mathcal{S}}_{k+1}$. If the candidate models are nested, $\widetilde{\mathcal{S}}_{k-1} = \mathcal{S}_{k-1}$ and $\widetilde{\mathcal{S}}_k = \mathcal{S}_k$.

Under H_0, $\mathcal{S}_{k_0} \subseteq \widetilde{\mathcal{S}}_{k-1} \subset \mathcal{S}_k \subset \widetilde{\mathcal{S}}_{k+1}$, and then asymptotically

$$A_{k,1} = 2\{\log L(\hat{\theta}(\widetilde{\mathcal{S}}_{k+1})) - \log L(\hat{\theta}(\mathcal{S}_k))\} \sim \chi^2_1,$$

$$A_{k,2} = 2\{\log L(\hat{\theta}(\mathcal{S}_k)) - \log L(\hat{\theta}(\widetilde{\mathcal{S}}_{k-1}))\} \sim \chi^2_1,$$

where χ^2_1 denotes the χ^2 distribution with one degree of freedom. Furthermore, by Cochran's theorem (Rao, 1973), $A_{k,1}$ and $A_{k,2}$ are independent because $A_{k,1}+A_{k,2} = 2\{\log L(\hat{\theta}(\widetilde{\mathcal{S}}_{k+1})) - \log L(\hat{\theta}(\widetilde{\mathcal{S}}_{k-1}))\} \sim \chi^2_2$ asymptotically.

Under H_1, similarly, if $\kappa(\text{GIC}_\alpha) = k_0$, asymptotically $A_{k_0,1} \sim \chi^2_1$ and $A_{k_0,2} \sim \chi^2_{1,v}$ are independent, where $\chi^2_{1,v}$ denotes the χ^2 distribution with one degree of freedom and with non-centrality parameter v. It is seen that $v = E\{A_{k_0,2}|k = k_0\} - 1$. So v increases with sample size, n.

3.3.2 Probability of Selecting a Model

Provided $\alpha > 0$, there is asymptotically zero probability of underfitting. In this section, we obtain upper bounds for the probability of selecting the correct model and also for selecting an overparameterized model.

Let $\kappa(\text{GIC}_\alpha)$ be the model size selected by GIC_α. From Proposition 3.1, the probability of selecting the model $\mathcal{S}_k$ by GIC_α is

$$\Pr\{\kappa(\text{GIC}_\alpha) = k\} = \Pr\{\alpha_{k,1} \le \alpha \le \alpha_{k,2}\}.$$

Under the null hypothesis H_0, $A_{k1} \le \alpha_{k,1} \le \alpha_{k,2} \le A_{k2}$, and $A_{k1} \sim \chi^2_1$ and $A_{k2} \sim \chi^2_1$ are independent. Hence the probability of selecting an overfitted model by GIC_α is asymptotically bounded by,

$$\begin{aligned} \Pr\{\kappa(\text{GIC}_\alpha) = k \mid k > k_0\} &\le \Pr\{A_{k1} \le \alpha \le A_{k2}\} \\ &= \Pr\{\chi^2_1 \le \alpha\}(1 - \Pr\{\chi^2_1 \le \alpha\}) \\ &= p. \end{aligned} \tag{3.2}$$

Similarly, the probability of selecting the true model by GIC_α is asymptotically bounded by

$$\begin{aligned} \Pr\{\kappa(\text{GIC}_\alpha) = k_0\} &\leq \Pr\{A_{k_0 1} \leq \alpha \leq A_{k_0 2}\} \\ &= \Pr\{\chi^2_1 \leq \alpha\}(1 - \Pr\{\chi^2_{1,v} \leq \alpha\}) \\ &= p_0. \end{aligned} \tag{3.3}$$

The upper bound, p, of the probability of selecting an overfitted model by GIC_α, is plotted in Figure 3.1.

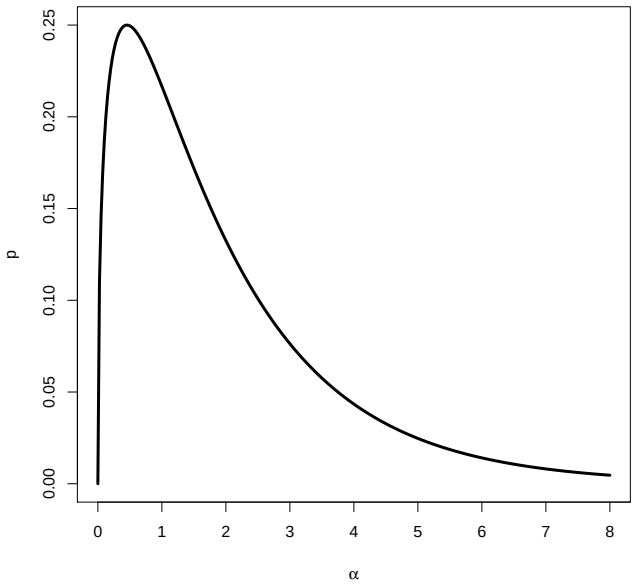

Figure 3.1: Upper bound probability, p, of selecting an overfitted model. The AIC corresponds to $\alpha = 2$ in which case the upper bound of the probability of selecting an overfitted model is about 13%. The maximum, $p = 0.25$, occurs at $\alpha = 0.455$.

The upper bound probabilities, p_0, from (3.3) are shown in Figure 3.2 for $v = 5, 10, 20, 40$. As v, or equivalently n, increases, p_0 increases.

The upper bound defined in Shao and Rao (2000, p. 217) may be written in our notation as the upper bound of the probability, $\Pr\{\kappa(\text{GIC}_\alpha) > k_0\}$.

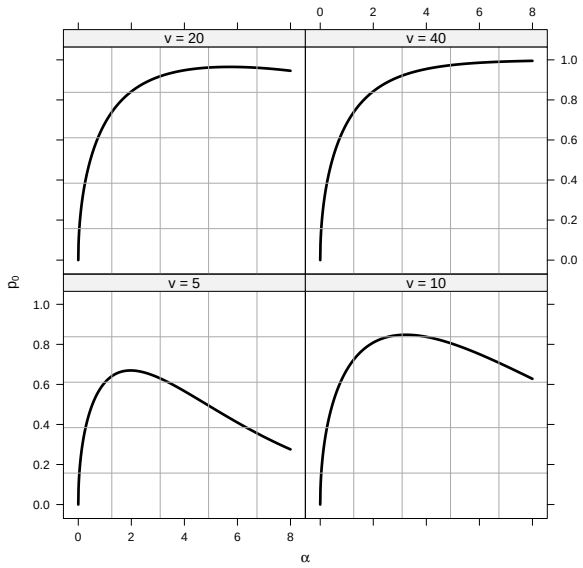

Figure 3.2: Upper bound probability, p_0, of selecting the true model with $v = 5, 10, 20, 40$.

3.3.3 Procedure

From (3.2) and Figure 3.1, the maximum probability of selecting an overfitted model occurs when $p = 0.25$ and $\alpha = 0.455$. Moreover, from (3.3) and Figure 3.2, if $\alpha \leq 0.455$, $p_0 \leq \Pr\{\chi_1^2 \leq \alpha\} \leq 0.5$. We may stipulate that the probability of selecting a true model is greater than 0.5 and take $\alpha > 0.455$. Then from (3.2), with $0 \leq p \leq 0.25$, $\Pr\{\chi_1^2 \leq \alpha\} = (1 + \sqrt{1 - 4p})/2$. Let $\tilde{p} = (1 + \sqrt{1 - 4p})/2$. Setting α_p,

$$\alpha_p = \chi_1^2(\tilde{p}), \tag{3.4}$$

which is the quantile for a χ_1^2 distribution, that is, $\Pr\{\chi_1^2 \leq \alpha_p\} = \tilde{p}$, defines the new procedure for selecting α. This procedure is denoted by GIC. In most cases, we may select $p \in [0.01, 0.1]$. In the next section, we will show that to achieve consistency we need $p \to 0$ as $n \to 0$. In practice, it is helpful to choose p small for larger n and also when the model space is large and most independent variables are not needed.

3.3.4 Consistent Procedure

From (3.3), the probability of selecting the true model by the GIC,

$$\Pr\{\kappa(\text{GIC}) = k_0\} \leq \Pr\{\chi^2_1 \leq \alpha_p\} = (1 + \sqrt{1 - 4p})/2.$$

So for a given constant p, the GIC, procedure (3.4), is inconsistent. Let

$$p_n = \Pr\{\chi^2_1 \leq \log n\}(1 - \Pr\{\chi^2_1 \leq \log n\}), \tag{3.5}$$

be the overfitting level instead of a fixed p, then $\alpha_p = \log n$, and the GIC is the same as the usual BIC (Schwarz, 1978) and is consistent.

We may set the overfitting level less than a given level p. Setting,

$$p_{1,n} = \min\{p, 1/\sqrt{n}\}, \tag{3.6}$$

and

$$p_{2,n} = \min\{p, p_n\}. \tag{3.7}$$

Then $n \leq 1/p^2$ and we obtain, $p_{1,n} = p$, for the upper level as in Shao and Rao (2000). When $n > e^a$, where a is the $1-p$ quantile of χ^2_1, $p_{2,n} = p_n$, the level for the BIC . Setting $p = p_{i,n}$, $i = 1$ and 2, the GIC is consistent.

There are other settings for the overfitting level that provide asymptotic consistency. We propose a setting of overfitting level such that the level is close to a given level p if the sample size is not large and other letting the level approach p_n in (3.5) as n gets larger. Whether the sample size is large or not is related to the ratio of n and K, the full model size. Roughly if $n/K > r_0$, where $r_0 \geq 2$, the sample size would be large. Let $c = 1/(1 + e^{-2(n/K - r_0)})$, and define the third rule,

$$p_{3,n} = (1 - c)p + c\{n/(n + 50)\}p_{2,n}. \tag{3.8}$$

Then GIC is consistent because $p_{3,n} - p_n \to 0$ as $n \to \infty$.

The three significant levels $p_{i,n}$, $i = 1,\ 2$ and 3 are plotted in Figure 3.3. It is seen that as n increases, $p_{2,n}$ and $p_{3,n}$ decrease at a faster rate than $p_{1,n}$. When n/K is smaller, $p_{3,n}$ is close to the upper level p. When n/K is larger, $p_{3,n}$ approaches $p_{2,n}$.

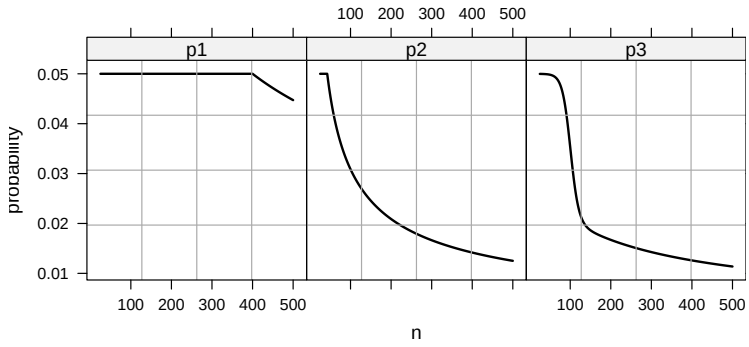

Figure 3.3: Three overfitting levels: $p_{1,n}$ (p1), $p_{2,n}$ (p2), $p_{3,n}$ (p3).

3.4 Simulations

3.4.1 Linear Regression with Overfitting Level $p = 0.01$

Consider the linear regression with $K = 5$ and $n = 40$,

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + e_i, i = 1, \ldots, 40,$$

where e_i are independent and identically distributed as $N(0, 1)$, $x_{i,1} = 1$, and $x_{i,k}$, $k = 2, \ldots, 5$, are specified in Shao (1993). The true values of β are shown in Table 3.1. We compare the performance of AIC, BIC, and GIC by measuring the percentage number of underfit, overfit and correct models, and the model error, $\| X\beta - X(\mathcal{S})\hat{\beta}(\mathcal{S}) \|^2$. The overfitting level $p = 0.01$ is used for GIC . We simulated 10^4 times for each parameter setting. The simulation results are shown in Table 3.1. It is seen that the GIC outperforms the AIC and BIC.

Table 3.1: Percentage number of underfitted models (u), correct models (c), and overfitted models (o), and true model error from 10^4 simulations for each parameter setting.

true β	procedure	u	c	o	model error
$(2, 0, 0, 4, 0)$	AIC	0	57	43	3.82
	BIC	0	82	18	2.98
	GIC	0	96	4	2.31
$(2, 0, 0, 4, 8)$	AIC	0	68	32	4.21
	BIC	0	87	13	3.67
	GIC	0	97	3	3.23

3.4.2 Comparison of Four Rules

There are four rules of setting the overfitting level p in our GIC procedure defined in (3.4). Using a constant, $p \in [0.01, 0.1]$, does not produce a consistent or asymptotically correct model choice but may nevertheless be useful in some applications. Some consistent model selection rules, $p_{1,n}$, $p_{2,n}$ and $p_{3,n}$, were given in (3.6), (3.7) and (3.8) respectively.

To compare the performance of these rules, we consider the linear model, as in Tibshirani (1996), $y = X\beta + e$, with $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$. The components of X and e are standard normal. The correlation between X_i and X_j is $\rho^{|i-j|}$ with $\rho = 0.5$. The true model size $k_0 = 3$. The full model size $K = 8$. Let $\sigma = 1$, and $n = 20, 60, 100$.

For each parameter setting, 10^4 simulations were done. The total number of the underfitted models, the true models, and the overfitted models selected by the GIC with the rules, $p_{1,n}$, $p_{2,n}$ and $p_{3,n}$ and a fourth rule, simply taking $p = 0.05$, was determined and the percentages are shown in Table 3.2. The standard deviation of each percentage can be calculated by the usual formula for proportions. The maximum standard deviation is 0.0032. The $r_0 = 5$ in the $p_{3,n}$-rule. For $n = 20$, the performance is the same with all four rules but $p_{3,n}$ outperforms others when $n = 60$ and $n = 100$.

Table 3.2: Percentage number of underfitted models (u), correct models (c), and overfitted models (o_k: k more variables) from 10^4 simulations. Comparison of $p = 0.05$ and rules $p_{1,n}$, $p_{2,n}$ and $p_{3,n}$ in eqns. (3.6), (3.7) and (3.8).

n	GIC	u	c	o_1	o_2	o_3	o_4
20	p	1	62	25	9	2	1
	$p_{1,n}$	1	62	25	9	2	1
	$p_{2,n}$	1	62	25	9	2	1
	$p_{3,n}$	1	62	25	9	2	1
60	p	0	74	21	4	0	0
	$p_{1,n}$	0	74	21	4	0	0
	$p_{2,n}$	0	77	19	3	0	0
	$p_{3,n}$	0	87	12	1	0	0
100	p	0	76	20	3	0	0
	$p_{1,n}$	0	76	20	3	0	0
	$p_{2,n}$	0	84	14	2	0	0
	$p_{3,n}$	0	89	10	1	0	0

3.4.3 Subset Autoregression

The AR (1) model, $z_t = \mu + \phi(z_{t-1} - \mu) + a_t$, $t = 1, \ldots, n$, where a_t is assumed independent normal with mean zero and variance σ_a^2. The model error, ME, was computed for best subset selection with up to $K = 10$ lags for subset autoregressions of the form, $z_t = \mu + \phi_{i_1}(z_{t-i_1} - \mu) + \ldots + \phi_{i_p}(z_{t-i_p} - \mu) + a_t$, where $i_1, \ldots, i_p \in \{1, \ldots, 10\}$ using AIC , BIC and GIC. With the GIC the overfitting level was set to $p = 0.01$. Denote the resulting estimates by $\hat{\phi}_{i_1}, \ldots, \hat{\phi}_{i_p}$. Then the corresponding observed model error may be written

$$\text{ME}\ (\hat{\varphi}) = (\hat{\varphi} - \varphi)' \mathcal{I}^{-1} (\hat{\varphi} - \varphi)/n,$$

where $\hat{\varphi} = (\hat{\varphi}_1, \ldots, \hat{\varphi}_{10})'$, $\varphi = (\phi, 0, \ldots, 0)'$,

$$\hat{\varphi}_i = \begin{cases} \hat{\phi}_i & i \in i_1, \ldots, i_p \\ 0 & \text{otherwise} \end{cases} \qquad \varphi_i = \begin{cases} \phi & i = 1 \\ 0 & \text{otherwise} \end{cases}$$

and $\mathcal{I} = \frac{1}{1-\phi^2}(\phi^{|i-j|})_{10\times 10}$ is the Fisher information matrix. The relative model error is the ratio ME $(\hat{\varphi})$/ME $(\hat{\phi})$, where $\hat{\phi}$ denotes the estimates in the full AR (K) model. For each parameter combination, 10^4 simulations were done. In Figure 3.4, the relative model errors are shown for the BIC and BIC_q. The AIC was omitted because the model error was very large. Only $\phi = 0, 0.3, 0.6, 0.9$ are shown since the results for other values are similar.

3.5 Illustrative Applications

3.5.1 South Africa Heart Disease Data

The data are described in Hastie et al. (2009, §4.4.2). The aim of the study was to establish the importance of ischemic heart disease risk factors. The response is a binary variable indicating the presence or absence of disease in 462 South African men. There are 9 predictors. The logistic regression model is used to fit the dataset. The AIC, BIC, and GIC with $p = 0.01$ select the same model that has the five variables `tobacco`, `ldl`, `famhist`, `typea` and `age`. Each of the selected variables is significant. If $p = 0.20$, the GIC selects a model with one more variable than the above, but the extra variable is not statistically significant at 10%. In this case, as expected, the selected model would be overfitting because $p = 0.20$ is set too high.

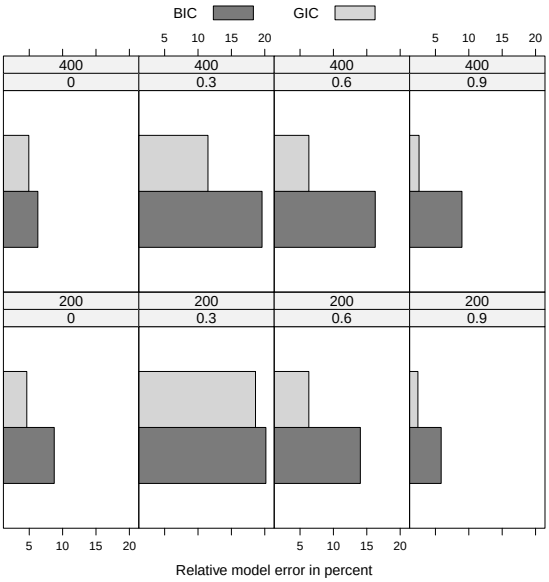

Figure 3.4: Relative model error in percent for AR(1) with $K = 10$ for series lengths $n = 200, 400$ and parameter setting $\phi = 0, 0.3, 0.6, 0.9$. GIC selection with $p = 0.01$.

3.5.2 Lynx Time Series

The lynx time series is derived from annual lynx population estimates from 1821−1934 in Canada and is discussed by Tong (1977) as well as many other researchers in time series. Tong (1977) fit a subset autoregression using the BIC and obtained the model with lags $1, 2, 4, 10, 11$. Using our package (McLeod et al., 2010) with $K = 15$ and with $p = 0.01$ for the GIC, we obtained the results shown in Table 3.3. Model diagnostic checks, including the portmanteau test and residual autocorrelation plot, indicate that the more parsimonious model that was selected using the GIC is adequate.

Table 3.3: Lags in subset autoregression selected by various information criterion

AIC	1	2	3	4	9	10	11
BIC	1	2	4	10	11		
GIC	1	2	9	12			

3.6 Conclusions

The generalized information criterion, GIC_α, includes a penalty parameter α. The performance relies on the choice of the α. The approximate efficiency (Shibata,

1984), bootstrap (Rao, 1999), and hypothesis testing (Shao and Rao, 2000) were introduced to choose the penalty parameter. We proposed a procedure by controlling the overfitting level, p. The procedure is consistent by controlling the overfitting level to be closer to zero. This method is computationally efficient as well as producing better model selection.

The essential difference between our method and that proposed by Shao and Rao (2000) is that we choose the penalty parameter α by controlling the probability of selecting an overfitting model, $\Pr\{\kappa(\text{GIC}_\alpha) = k \mid k > k_0\}$, instead of the overfitting probability, $\Pr\{\kappa(\text{GIC}_\alpha) > k_0\}$, considered by Shao and Rao (2000). Our approach is simpler and has been implemented in our software packages for generalized linear models (McLeod and Xu, 2010) and subset autoregression (McLeod et al., 2010). Some illustrative applications are given in §3.5 and more are available in the documentation in our software packages.

Chapter 4

FAMILY OF BAYESIAN INFORMATION CRITERIA

The family of Bayesian information criteria using the Bernoulli prior, BIC_q, with parameter $q \in (0,1)$, is discussed. The BIC_q is an effective criterion for many types of Bayesian model selection problems. We establish some new theorems that elucidate the behavior of the BIC_q and suggest its suitability for large model spaces as well as many other kinds of model selection problems. Simulation studies are presented that demonstrate that the BIC_q is more effective than the usual BIC in many situations. Several interesting applications are also examined. The BIC_q is implemented in our packages for the subset selection in the generalized linear model (McLeod and Xu, 2010) as well as for autoregressive models (McLeod et al., 2010). Scripts are provided in the vignettes accompanying these packages for reproducing all figures and tables in this chapter.

4.1 Introduction

Let $y = (y_1, ..., y_n)$ be a vector of responses and $X = (X_1, ..., X_d)$ be an $n \times d$ matrix of inputs. Let $\mathcal{S}_k = \{s_1, ..., s_k\}$ be a subset of $\{1, 2, ..., d\}$, which represents a class of models with size k. The model is specified by a distribution function $f_{\theta(\mathcal{S}_k)}(y|X(\mathcal{S}_k))$, where $\theta(\mathcal{S}_k)$ is a vector of the parameters, and $X(\mathcal{S}_k)$ denotes the matrix formed by selecting the columns corresponding to $\mathcal{S}_k$ from X. After the data is available, let $L(\theta(\mathcal{S}_k)) = f_{\theta(\mathcal{S}_k)}(y|X(\mathcal{S}_k))$ be the likelihood function and $\hat{\theta}(\mathcal{S}_k)$ the maximum likelihood estimate.

We consider model selection using Bayesian information criterion with a Bernoulli prior. A general family of Bayesian information criteria (Hansen and Yu, 2001; Rissanen, 2007) may be written,

$$-2\log L(\hat{\theta}(\mathcal{S}_k)) + k\log n - 2\log p(\mathcal{S}_k),$$

where $p(\mathcal{S}_k)$ is a prior probability of the model determined by $\mathcal{S}_k$ and $k = \kappa(\mathcal{S}_k)$ is the model size. Assuming that $p(\mathcal{S}_k)$ is a constant, Schwarz (1978) obtained the widely used (Hastie et al., 2009, §7.7) BIC criterion,

$$\text{BIC} = -2 \log L(\hat{\theta}(\mathcal{S}_k)) + k \log n.$$

(George and Foster, 2000, eqn (6)) suggested using a Bernoulli prior with parameter $q \in (0, 1)$. In this formulation q is the probability that each parameter appears in the model. This implies that the prior may be written as $p(\mathcal{S}_k) = q^k(1-q)^{K-k}$, where K is the maximum model size. Hence, dropping the constant term involving K,

$$\text{BIC}_q = -2 \log L(\hat{\theta}(\mathcal{S}_k)) + k \log n - 2k \log[q/(1-q)].$$

Computationally, it is convenient to denote the value of BIC_q for a specified value of q, $q = q_0$, by $\text{BIC}(q = q_0)$ and to extend the definition so that $q = 0$ corresponds to the null model with no parameters selected and $q = 1$ corresponds to the full model.[1]

When $q = 0.5$, the BIC_q is same as BIC and more generally the BIC_q shares the consistency property of the BIC, that is, provided the correct model is included in the possible candidate models, and q is held fixed as $n \longrightarrow \infty$, the correct model will be chosen with probability one (Shao, 1997; Yang, 2005).

In the next sections, we will show that the BIC_q provides a more general and flexible Bayes information criterion than the extended Bayesian information criterion, BIC_γ (Chen and Chen, 2008).

4.2 Properties

4.2.1 BIC_q More General Than BIC_γ

The extended Bayesian information criterion suggested by Chen and Chen (2008) may be written as

$$\text{BIC}_\gamma = -2\text{log}L(\hat{\theta}(\mathcal{S}_k)) + k \log n + 2\gamma \log \text{C}(K, k),$$

1. Note that in variable selection problems in regression, the intercept term is usually included in all models, so it is not counted as a parameter and in this case, the null model corresponds to the model with only an intercept term.

where $0 \leq \gamma \leq 1$ and $\mathrm{C}(K, k)$ denotes the number of combinations, K choose k. This criterion was derived for large model spaces since in this case, it may be thought that all models having the same size, k, should be equally likely. Specifically, Chen and Chen (2008) suggested the prior, $p(\mathcal{S}_k) \propto [\mathrm{C}(K, k)]^{-\gamma}$. When $\gamma = 1$, all models having the same size are assumed equally likely and when $\gamma = 0$, it reduces to the usual BIC. Regardless of the value of γ, the prior specifies that average model size is $E\{\kappa(\mathcal{S}_k)\} = K/2$. The parameter γ in the BIC_γ was introduced in an ad hoc fashion and has no useful interpretation unlike q in the BIC_q.

On the other hand, for the BIC_q, $E\{\kappa(\mathcal{S}_k)\} = qK$. So for the large model space problem where K is large and not many parameters are expected in the final model, $q < 1/2$, seems more reasonable than the assumption, implicit in the BIC_γ, that the average number of parameters is $K/2$. By varying q, the BIC_q is suitable for a wide range of statistical problems such as in prediction or smoothing. In Theorem 4.3 we show that BIC_q provides a more general criterion than the BIC_γ.

Let $\kappa(\mathrm{BIC})$, $\kappa(\mathrm{BIC}_q)$, and $\kappa(\mathrm{BIC}_\gamma)$ denote the size of the model selected by the BIC, BIC_q, or BIC_γ respectively.

The following lemma is useful for the proofs of Theorems.

Lemma 4.1. *Let $\mathrm{C}_i(k)$, $i = 1, 2$, be two criteria. Let $\kappa(\mathrm{C}_i)$ denote the model size selected by $\mathrm{C}_i(k)$. Define $\Delta(k) = \mathrm{C}_2(k) - \mathrm{C}_1(k)$. Then $\kappa(\mathrm{C}_1) \geq \kappa(\mathrm{C}_2)$ if $\Delta(k)$ increases and $\kappa(\mathrm{C}_1) \leq \kappa(\mathrm{C}_2)$ if $\Delta(k)$ decreases.*

Proof. We prove the first part. The second part is only the converse of the first. Assume that $\Delta(k)$ increases. Let $k_i = \kappa(\mathrm{C}_i)$. Then $\mathrm{C}_i(k) \geq \mathrm{C}_i(k_i)$ for all k. For $k > k_1$, we have $\Delta(k) > \Delta(k_1)$ and $\mathrm{C}_1(k) \geq \mathrm{C}_1(k_1)$. Thus

$$\mathrm{C}_2(k) = \mathrm{C}_1(k) + \Delta(k) > \mathrm{C}_1(k_1) + \Delta(k_1) = \mathrm{C}_2(k_1).$$

So k_2 cannot be greater than k_1, that is, $k_2 \leq k_1$. □

Let $q_1 < q_2$ and $k_i = \kappa(\mathrm{BIC}_{q_i})$. Then

$$\Delta(k) = \mathrm{BIC}_{q_2} - \mathrm{BIC}_{q_1} = 2k \log q_1(1 - q_2)/\{q_2(1 - q_1)\},$$

decreases. From Lemma 4.1 we have $k_1 \leq k_2$. Hence, increasing q causes the number of parameters selected to increase or to stay the same.

Theorem 4.1. *For each $\gamma \in [0,1]$, there exists $q = q_\gamma$ such that $\kappa(\mathrm{BIC}_q) = \kappa(\mathrm{BIC}_\gamma)$. Let $k_\gamma = \kappa(\mathrm{BIC}_\gamma)$. Then*

$$q_\gamma = \begin{cases} 1/[1 + \{(K - k_\gamma)/(k_\gamma + 1)\}^\gamma], & k_\gamma < K \\ 1/(1 + 1/K^\gamma), & k_\gamma = K. \end{cases}$$

Proof. Let $0 \le \gamma \le 1$, $k_\gamma = \kappa(\mathrm{BIC}_\gamma)$, and

$$\Delta(k) = \mathrm{BIC}_q(k) - \mathrm{BIC}_\gamma(k) = -2\log\{q^k(1-q)^{-k}\mathrm{C}(K,k)^\gamma\}.$$

If $k_\gamma < K$, q_γ is determined from the equation,

$$\Delta(k+1) - \Delta(k) = -2\log\{(K-k)/(k+1)\}^\gamma q/(1-q) = 0,$$

by letting $k = k_\gamma$, that is,

$$q_\gamma = [1 + \{(K - \kappa_\gamma)/(\kappa_\gamma + 1)\}^\gamma]^{-1}.$$

If $\kappa_\gamma = K$, let $q_\gamma = \{1/(1 + K^{-\gamma})\}^{-1}$.

Let $q = q_\gamma$. Then k_γ is the minimum point of $\Delta(k)$. The $\Delta(k)$ decreases if $k < k_\gamma$ and increases if $k > k_\gamma$. Hence from Lemma 4.1, the $\mathrm{BIC}_q(k)$ has the minimum at k_γ. That is, $\kappa(\mathrm{BIC}_q) = \kappa(\mathrm{BIC}_\gamma) = k_\gamma$. □

Theorem 4.1 shows that the model selected by the BIC_γ can also be selected using the BIC_q.

Theorem 4.2. *Let $0 \le \gamma_1 < \gamma_2 \le 1$, and assume that BIC_γ has a unique minimum. Then if $\kappa(\mathrm{BIC}) < K/2$, $\kappa(\mathrm{BIC}_{\gamma_2}) \le \kappa(\mathrm{BIC}_{\gamma_1}) \le \kappa(\mathrm{BIC})$, otherwise if $\kappa(\mathrm{BIC}) > K/2$, $\kappa(\mathrm{BIC}_{\gamma_2}) \ge \kappa(\mathrm{BIC}_{\gamma_1}) \ge \kappa(\mathrm{BIC})$.*

Proof. Let $k_i = \kappa(\mathrm{BIC}_{\gamma_i})$ and $K_0 = K/2$. Assume that $\gamma_1 < \gamma_2$. Then

$$\Delta(k) = \mathrm{BIC}_{\gamma_2}(k) - \mathrm{BIC}_{\gamma_1}(k) = 2(\gamma_2 - \gamma_1)\log \mathrm{C}(K,k),$$

increases for $k < K_0$ and decreases for $k > K_0$. Since $\mathrm{BIC}_{\gamma_i}(k)$ has a unique minimum, the $\mathrm{BIC}_{\gamma_i}(k)$ decreases for $k < k_i$ and increases for $k > k_i$.

Suppose that $k_1 < K_0$. Let $k_1 < k < K_0$. Then $\mathrm{BIC}_{\gamma_1}(k_1) < \mathrm{BIC}_{\gamma_1}(k) < \mathrm{BIC}_{\gamma_1}(K_0)$, and $\Delta(k_1) < \Delta(k) < \Delta(K_0)$. From $\mathrm{BIC}_{\gamma_2}(k) = \mathrm{BIC}_{\gamma_1}(k) + \Delta(k)$,

$\text{BIC}_{\gamma_2}(k_1) < \text{BIC}_{\gamma_2}(k) < \text{BIC}_{\gamma_2}(K_0)$. Hence the minimum of $\text{BIC}_{\gamma_2}(k)$ should be no more than k_1, that is, $k_2 \leq k_1 < K_0$.

Similarly, if $k_1 > K_0$, for $K_0 < k < k_1$, we have $\text{BIC}_{\gamma_1}(K_0) > \text{BIC}_{\gamma_1}(k) > \text{BIC}_{\gamma_1}(k_1)$ and $\Delta(K_0) > \Delta(k) > \Delta(k_1)$, and then $\text{BIC}_{\gamma_2}(K_0) > \text{BIC}_{\gamma_2}(k) > \text{BIC}_{\gamma_2}(k_1)$. Hence the minimum of $\text{BIC}_{\gamma_2}(k)$ should be no less than k_1, that is, $k_2 \geq k_1 > K_0$.

Let $\gamma_0 = 0$. $\kappa(\text{BIC}_{\gamma_0}) = \kappa(\text{BIC})$. Hence if $\kappa(\text{BIC}_{\gamma_0}) < K_0$, $\kappa(\text{BIC}_{\gamma_2}) \leq \kappa(\text{BIC}_{\gamma_1}) \leq \kappa(\text{BIC})$. And if $\kappa(\text{BIC}_{\gamma_0}) > K_0$, $\kappa(\text{BIC}_{\gamma_2}) \geq \kappa(\text{BIC}_{\gamma_1}) \geq \kappa(\text{BIC})$. □

Theorem 4.2 shows that as γ increases, the number of parameters selected in the model may increase or decrease depending on K and the number of parameters selected using the BIC.

Theorem 4.3. BIC_q *provides a more general criterion than* BIC_γ.

Proof. From Theorem 4.2, the selected model size by BIC_γ is either less or greater than that by BIC. Assume that $\kappa(\text{BIC}) < K/2$. Then for each $\gamma \in [0, 1]$, $\kappa(\text{BIC}_\gamma) \leq \kappa(\text{BIC}) < K/2$. If q is close to 1, $\kappa(\text{BIC}_q)$ approaches K. So there may exist $q \in (0, 1)$ such that for all $\gamma \in [0, 1]$, $\kappa(\text{BIC}_q) \neq \kappa(\text{BIC}_\gamma)$. Hence there may be some cases where the BIC_γ cannot select a model that can be selected using the BIC_q. And from Theorem 4.1, any model selected using the BIC_γ may also be selected using the BIC_q. □

4.2.2 Tuning Parameter

The performance of the BIC_q relies on the tuning parameter q. The Bernoulli prior specifies that the average model size is $E\{\kappa(\mathcal{S}_k)\} = qK$. For many screening or subset selection problems, especially when K is large, we have found that $q = 0.25$ works well. For most problems this produces a model that is more parsimonious than a model selected using the BIC.

As is common practice in time series model building (Box et al., 2005, §1.3.2), an iterative model building approach involving initial model selection followed by diagnostic checking, refitting and rechecking is recommended. Adopting this approach with the BIC_q, we may start with $q = 0.25$ to determine an initial model. After suitable diagnostic checks, if it is found that the model is inadequate, a larger value of q may be tried. Or if a more parsimonious model is required, we may refit using a smaller value of q.

Alternatively, as suggested by George and Foster (2000), cross-validation could be used. Another possibility would be to use bootstrapping to choose the value of q to minimize the prediction error. Both these approaches are non-Bayesian as well as more laborious than needed in some cases. If the cross-validation or bootstrapping approach is used, it would seem more natural to consider the generalized Akaike information criterion (Bhansali and Downham, 1977; Akaike, 1979),

$$\text{AIC}_\alpha(k) = -2\log L(\hat{\theta}(\mathcal{S}_k)) + \alpha k,$$

where α is a tuning parameter. It may be shown that AIC_α and BIC_q are equivalent[2] in the sense that taking $\alpha = \log n - 2\log q/(1-q)$, the AIC_α will then select the same model as the BIC_q and similarly for any model chosen using the AIC_α, taking $q = 1/(e^{\alpha/2}/\sqrt{n}+1)$, will result in the same model being chosen using the BIC_q.

In some problems, such as spectral density function estimation by autoregression, the cost function may not be evident and so neither cross-validation or bootstrapping is likely to be useful. The iterative model building approach may be adequate in many situations.

4.3 Simulation Experiments

The simulations reported in this section suggest that often the BIC_q with $q = 0.25$ will outperform the BIC.

4.3.1 Linear Regression

We consider the following linear model

$$y_i = x_i^T(\mathcal{S})\beta(\mathcal{S}) + \epsilon_i, i = 1, \ldots, n,$$

where $\mathcal{S}$ is a subset of $\{1, ..., K\}$, and ϵ_i are independent and identically distributed as $N(0, \sigma^2)$. The K covariates are generated from the multinormal distribution, $N(\mathbf{0}, \Sigma)$ with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.2$ for $i \neq j$. The model was examined in Chen and Chen (2008). Let $b_0 = (0.7, 0.9, 0.4, 0.3, 1.0, 0.2, 0.2, 0.1)$ be a vector of length 8. The

2. Note that if q is allowed to depend on n, the BIC_q would lose the consistency property.

following two models are examined: (1) $\beta(\mathcal{S}) = b_0$, and (2) $\beta(\mathcal{S}) = (b_0, b_0)$. In this example $K = 20$, $n = 200$ and $\sigma = 0.2$.

We compare the performance of AIC, BIC, and BIC_q with $q = 0.25$ by measuring the number of underfitted, overfitted and correct models, and the model error (ME), $\| X\beta - X(\mathcal{S})\hat{\beta}(\mathcal{S}) \|^2$. We simulated 100 times for each parameter setting. The simulation results are shown in Table 4.1. It is seen that the BIC_q outperforms the AIC and BIC.

Table 4.1: The number of underfit, overfit and correct models, and the model error

True $\beta(\mathcal{S})$	procedure	overfit	underfit	correct	ME
β_0	AIC	86	0	14	0.574
	BIC	21	0	79	0.382
	BIC_q	9	0	91	0.345
(β_0, β_0)	AIC	59	0	41	0.709
	BIC	11	0	89	0.637
	BIC_q	6	0	94	0.625

4.3.2 Subset Autoregression AR(1)

The model error, ME, was computed for best subset selection with up to $K = 10$ lags for subset autoregressions of the form, $z_t = \mu+\phi_{i_1}(z_{t-i_1}-\mu)+\ldots+\phi_{i_p}(z_{t-i_p}-\mu)+a_t$, where $i_1, \ldots, i_p \in \{1, \ldots, 10\}$ using AIC , BIC and BIC_q. For the BIC_q, $q = 0.25$. The underlying model was an AR (1) model, $z_t = \mu+\phi(z_{t-1}-\mu)+a_t$, $t = 1, \ldots, n$, where a_t is assumed independent normal with mean zero and variance σ_a^2. Denote the resulting estimates by $\hat{\phi}_{i_1}, \ldots, \hat{\phi}_{i_p}$. Then the corresponding observed model error may be written ME $(\hat{\varphi}) = (\hat{\varphi} - \varphi)'\mathcal{I}^{-1}(\hat{\varphi} - \varphi)/n$, where $\hat{\varphi} = (\hat{\varphi}_1, \ldots, \hat{\varphi}_{10})'$, $\varphi = (\phi, 0, \ldots, 0)'$,

$$\hat{\varphi}_i = \begin{cases} \hat{\phi}_i & i \in i_1, \ldots, i_p \\ 0 & \text{otherwise} \end{cases} \qquad \varphi_i = \begin{cases} \phi & i = 1 \\ 0 & \text{otherwise} \end{cases}$$

and $\mathcal{I} = \frac{1}{1-\phi^2}(\phi^{|i-j|})_{10\times 10}$ is the Fisher information matrix. The relative model error is the ratio ME $(\hat{\varphi})$/ME $(\hat{\phi})$, where $\hat{\phi}$ denotes the estimates in the full AR (K) model. For each parameter combination, 10^4 simulations were done. In Figure 4.1, the relative model errors are shown for the BIC and BIC_q. The AIC was omitted because it was very large. Only $\phi = 0, 0.3, 0.6, 0.9$ are shown since the results for other values are similar.

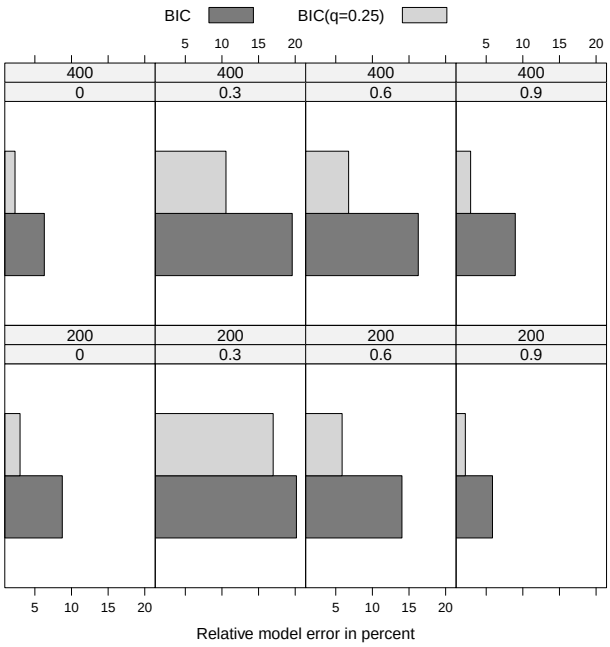

Figure 4.1: Relative model error in percent for AR(1) with $K = 10$ for series lengths $n = 200, 400$ and parameter setting $\phi = 0, 0.3, 0.6, 0.9$. BIC($q = 0.25$).

4.3.3 Subset Autoregression AR(4)

The autoregressive process of order 4, $z_t = \mu + \phi_1(z_{t-1} - \mu) \ldots + \phi_4(z_{t-4} - \mu) + a_t$, $t = 1, \ldots, n$, with $\phi_1 = 2.7607$, $\phi_2 = -3.8106$, $\phi_3 = 2.6535$, and $\phi_4 = -0.9238$ has been widely used as an example of a time series whose time series has two peaks in the spectral density that are close together (Percival and Walden, 1993, p. 42). In our simulations, it is assumed that a_t is Gaussian white noise with mean zero, unit variance and that $\mu = 0$. The method given in McLeod et al. (2010) was used to fit a subset autoregression with maximum order $K = 30$ with the BIC and BIC_q with $q = 0.25$. The empirical probability based on 10^4 simulations of including a parameter at lag k was determined. For $k = 1, 2, 3, 4$ it was found that this probability was exactly one. For $k > 4$ the lower this probability, the better the performance of the selection criterion. In the oracle case, corresponding to perfect selection, this probability is zero. From Figure 4.2, it is seen that the BIC_q with $q = 0.25$ decisively outperforms the BIC.

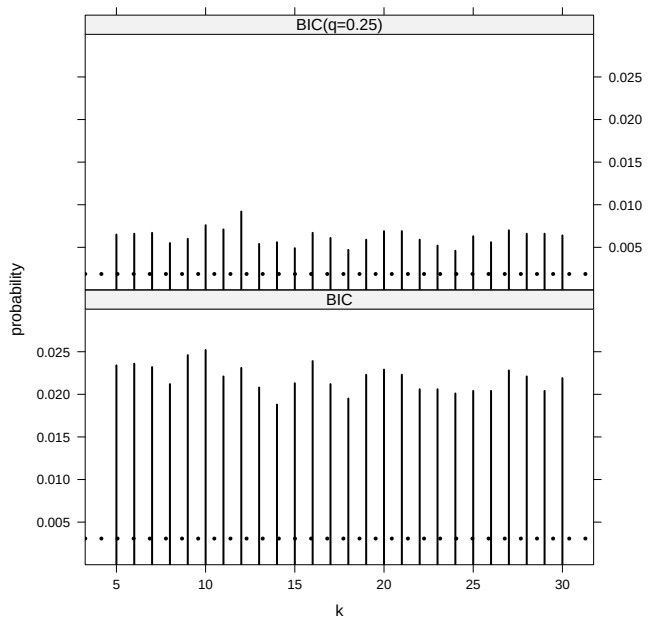

Figure 4.2: The empirical probability of including lag k in a subset autoregression with $K = 30$ based on 10^4 simulations of an AR(4) time series. The dotted line shows the conservative estimate of a 95% margin of error.

4.4 Illustrative Applications

The applications discussed also suggest than the BIC_q may be preferable in many situations.

4.4.1 Hospital Manpower Data

This dataset, taken from Myers (1990, Table 3.8), has 5 inputs x_1, x_2, x_3, x_4, and x_5. The AIC and BIC both select a model with three inputs. But one of these inputs is not even significant at the 5% level and it has a negative regression coefficient while a positive one was anticipated. Using the BIC_q with $q = 0.25$ results in a model with only two inputs and both of them are highly significant and have the correct signs. This model was also recommended by Myers (1990, p. 292). This example also provides an illustration of Theorem 4.3 since using the BIC_γ with all values of $\gamma \in [0, 1]$, a model with three or more inputs is always selected.

4.4.2 Monthly Sunspot Series $1749 - 1997$

This series, `sunspot.month` of length $n = 2988$, is included in the built-in datasets in R (R Development Core Team, 2010). Subset autoregressions were fit to this series using the BIC_q with $q = 0.5$ and $q = 0.25$ and the resulting estimates of the spectral density function are shown in Figure 4.3. The BIC_q with $q = 0.5$ is the BIC. The plot with $q = 0.25$ is smoother and is preferred over the more noisy plot with $q = 0.5$.

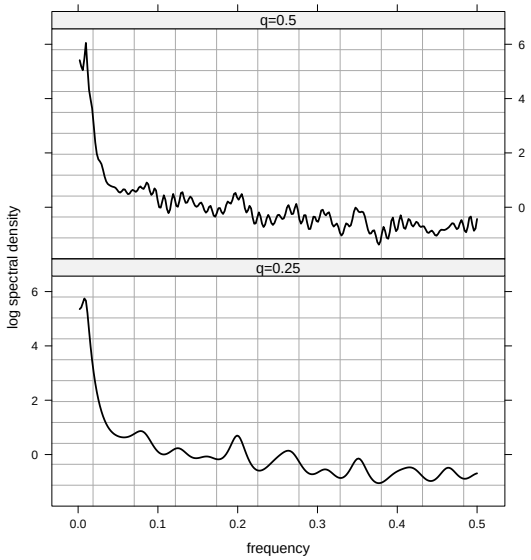

Figure 4.3: Estimated log spectral density function estimated by fitting a subset autoregression using BIC_q with $q = 0.5$ and $q = 0.25$.

4.4.3 Long Autoregressions

In time series prediction problems, model parsimony is a well-established principle (Granger and Jeon, 2004; McLeod, 1993) but as noted by Hastie et al. (2009, §7.7) the BIC may chose models which are too parsimonious in some applications. For example, often a sufficiently long autoregression is needed to capture the salient aspects of the time series in selecting the model order for autoregressive spectral density estimation and for estimating the inverse autocorrelations. In these problems we may use the BIC_q with $q > 1/2$.

In this order selection problem, we need to choose p in the $\text{AR}(p)$ model, $z_t = \zeta + \phi_1 z_{t-1} + ... + \phi_p z_{t-p} + a_t$, where $t = 1, \ldots, n$, a_t is assumed independent normal

with mean zero and constant variance, $\phi_1, \ldots, \phi_p$ are the autoregressive coefficients, and ζ is the intercept term. The autoregressive parameters are assumed to satisfy the condition for stationarity (Box et al., 2005, §3.1.3).

For autoregressive estimation of the spectral density function a sufficiently long autoregression is needed to capture the peaks in the spectral density with precision (Percival and Walden, 1993, Ch. 9). In these applications the usual BIC does not produce satisfactory estimates. Percival and Walden (1993, Ch. 9) use the final prediction error criterion (FPE) for choosing the order of the autoregression for spectral density estimation. The FPE is essentially equivalent to the AIC. Kay (1988, Ch 9.5) also uses the AIC for autoregressive spectral density estimation.

Similarly, when using autoregression to estimate the inverse autocorrelations, a sufficiently high order is needed (Chatfield, 1979, §4) and (Hipel and McLeod, 1994, §5.3.6).

In all of these applications, the BIC_q may be used in place of the AIC. In Section 5.5, we note that the BIC_q is capable of choosing any model that may be chosen using the AIC or even the more general generalized AIC.

In Table 4.2, we compare the model order p that is selected for some time series that exhibit periodicity. Each of the time series is available in R and the reader may wish to read the documentation supplied for more information. Several of the series are of particular interest. For autoregressive spectral density estimation of the `Willamette` series, Percival and Walden (1993, p. 520) recommended using either $p = 27$ or $p = 38$ for the logarithms of the series based on the final prediction error criterion. Cleveland (1972) used $p = 7$ or $p = 10$ for estimation of the inverse autocorrelations to `SeriesA`.

Table 4.2: The table shows p, the order selected for fitting an AR(p) to some time series with peak spectra of various lengths, n. The series `Willamette` and `SeriesA` are available in the R package `FitAR` (McLeod et al., 2010) and `lynx` and `sunspot.year` are included in the base distribution of R (R Development Core Team, 2010). The series `sunspot.year` are the mean annual sunspot numbers for the period $1700-1988$.

Name	n	AIC	BIC	q=0.75	q=0.8	q=0.85	q=0.9	q=0.95
Willamette	395	38	11	11	11	23	34	34
SeriesA	197	7	2	2	2	7	14	15
lynx	114	11	2	11	11	11	11	11
sunspot.year	289	9	9	9	9	9	22	24

4.5 Concluding Remarks

By using the tuning parameter $q \in (0, 1)$, the BIC_q provides a more flexible Bayesian information criterion than either the BIC or BIC_γ. The BIC_q may be used for large model spaces, prediction or smoothing problems. With $q < \frac{1}{2}$ more smoothing[3] is done than with the usual BIC. If less smoothing is desirable, $q > \frac{1}{2}$ may be used.

There are other approaches to Bayesian model selection that may be preferable in some situations (Robert, 2007, §7).

3. More smoothing corresponds to fewer estimated parameters.

Chapter 5

GIC FOR HIGH DIMENSIONAL MODEL SELECTION

Penalized maximum likelihood estimation was proposed for high dimensional variable selection. The performance of penalized likelihood method relies on the choice of regularization parameter. We consider the generalized information criterion for choosing the regularization parameter. We derive the conditions under which the criterion is overfitting, consistent, or underfitting. Our results are illustrated by simulation examples.

5.1 Introduction

In many high dimensional modelling problems, the number of variables is large but the number of significant variables is small. The traditional best subset selection procedure becomes infeasible for the high dimensional problem due to computational cost. Furthermore, the best subset selection is unstable with respect to a small change in the data (Breiman, 1996). Penalized maximum likelihood estimation has been suggested to automatically select significant variables (Tibshirani, 1996; Fan and Li, 2001). The penalized likelihood methods are computationally efficient, and with a proper choice of regularization parameter, achieve selection consistency or the oracle property (Fan and Li, 2001; Fan and Lv, 2010). The oracle property means that, asymptotically, the resulting statistical estimates have the same covariance matrix as when the correct variables are known a priori.

Penalized likelihood methods with various penalty functions have been developed, such as least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010). These methods produce a set of candidate models for different regularization parameters. With a proper choice of the regularization parameter, the resulting model will be consistent. Cross-validation (CV) is often used to choose the regularization parameter and if this is done correctly, consistent model

selection can be achieved (Shao, 1993, 1997). Since each regularization parameter corresponds to a model, choosing the proper regularization parameter is equivalent to selecting the best model from the set of candidate models. Hence, automatic model selection criteria, such as the BIC (Schwarz, 1978), are also applicable for determining the regularization parameter. For a linear model, Wang et al. (2007) analyzed asymptotic properties of the generalized CV and BIC for choosing the regularization parameter of the SCAD and showed that the BIC is consistent whereas the generalized CV yields overfitting.

The generalized information criterion (GIC) (Akaike, 1979; Nishii, 1984; Bhansali, 1986) includes a wide range of model selection criteria as a special case (Shao, 1997; Zhang, 2009). For a generalized linear model, Zhang et al. (2010) obtained the asymptotic properties of the GIC for choosing the regularization parameter of the non-concave penalized likelihood methods, such as SCAD and MCP. In this paper, we examine the asymptotic properties of the GIC for choosing the regularization parameter in a general case. The model is described by a family of probability distribution, which includes the generalized linear model. The penalty function for the penalized likelihood methods can be either non-concave, such as SCAD and MCP, or concave, as LASSO.

We derive the conditions under which the GIC is overfitting, consistent, or underfitting. When the penalized likelihood methods possess the oracle property, the GIC with both penalized and non-penalized maximum likelihood estimators have the same asymptotic properties. In this case, statistical inferences for the selected model are the same as if variables in the model were initially known. On the other hand, when the penalized likelihood method doesn't possess the oracle property, the BIC may not be consistent due to the bias between non-penalized and penalized estimators.

The asymptotic properties of the GIC with MLE have been widely investigated (Nishii, 1984; Sin and White, 1996; Shao, 1997; Yang, 2005). These authors considered the special case of the GIC with MLE and required that for consistency $\alpha \to \infty$ and $\alpha/n \to 0$. We obtain more general conditions for the penalized MLE under which the GIC is overfitting, consistent, or underfitting. And we show that in the special case of MLE, the condition $\alpha/n \to 0$ is sufficient but not necessary for consistency.

5.2 Penalized MLE Model Selection

5.2.1 Probability Models

Consider a family of probability distributions, $f(z;\theta)$, $z \in \Re^p$, indexed by parameters $\theta \in \Theta \subset \Re^d$. Let $\mathcal{S} = \{s_1, ..., s_k\}$ be a subset of $\{1, 2, ..., d\}$. Each subset $\mathcal{S}$ represents a class of probability models $\{f(z;\theta) : \theta \in \Theta(\mathcal{S})\}$, where $\Theta(\mathcal{S}) = \{\theta \in \Theta : \theta_i = 0, i \notin \mathcal{S}\}$. Let $\kappa(\mathcal{S})$ be the number of elements in $\mathcal{S}$, and $\theta(\mathcal{S}) \in \Theta(\mathcal{S})$. Let $z_1, ..., z_n$ be n observations that are independent and identically distributed with $Z_i \sim f(z;\theta_0)$, where $\theta_0 \in \Theta(\mathcal{S}_0)$, $\mathcal{S}_0$ represents a true model specified by

$$\theta_0 = \arg\max_{\theta\in\Theta} E\{\log f(Z;\theta)\}.$$

The observation $z_i \in \Re^p$ consists of both response and explanatory variables. Let $L_n(\theta) = \prod_i f(z_i;\theta)$, be the likelihood function and $l_n(\theta) = \log L_n(\theta)$. Let $\hat{\theta}_{\mathrm{M}}(\mathcal{S})$ denote the MLE of $\theta(\mathcal{S})$.

Assume that the true model is identifiable and its size $\kappa(\mathcal{S}_0) = k_0$ is finite. The true model is identifiable if for fixed n and each subset $\mathcal{S} \neq \mathcal{S}_0$ with $\kappa(\mathcal{S}) \leq \kappa(\mathcal{S}_0)$ and some $\Delta > 0$,

$$\max_{\theta\in\Theta(\mathcal{S}_0)} n^{-1} l_n(\theta) - \max_{\theta\in\Theta(\mathcal{S})} n^{-1} l_n(\theta) \geq \Delta. \tag{5.1}$$

5.2.2 Penalized MLE

The penalized MLE is obtained by maximizing the penalized likelihood (Fan and Li, 2001)

$$l_n(\theta) - n\sum_{j=1}^{d} p_\lambda(|\theta_j|), \tag{5.2}$$

where $p_\lambda(\cdot)$ is a penalty function indexed by the regularization parameter $\lambda \geq 0$. With a suitable penalty function, maximizing the penalized likelihood can simultaneously select variables and estimate their coefficients. The resulting penalized MLE may possess three properties: sparsity, unbiasedness and continuity (Fan and Li, 2001). With sparsity, the penalized MLE can achieve variable selection consistency. With both sparsity and unbiasedness, the penalized MLE may have the oracle property. With continuity, the penalized MLE is stable, that is, the estimates are not sensitive to small changes in the data.

The widely used penalties include ℓ_q penalties (Frank and Friedman, 1993) having ℓ_1 and ℓ_2 penalties as special cases and quadratic spline penalties such as SCAD and MCP (Fan and Li, 2001; Zhang, 2010). The quadratic spline penalties also include the ℓ_1 penalty. In the ℓ_q penalty family, only ℓ_1 penalty, used in LASSO (Tibshirani, 1996), produces a sparse estimator with the continuity property but it is biased. With the quadratic spline penalties and suitable regularization parameters, the penalized MLE has both sparsity and continuity as well as approximate unbiasedness (Fan and Lv, 2010; Zhang, 2010).

For each parameter λ, maximizing the penalized likelihood (5.2) produces one model. As λ varies in a range, we obtain a set of models, the candidate models. With a suitable penalty, the best model having the oracle property would asymptotically be included in this set.

5.2.3 Algorithms for Penalized MLE

For a linear model, $y = X\beta + \sigma e$, where X is an $n \times d$ matrix and $e \sim N(\mathbf{0}, \mathbf{1}_n)$. Maximizing the penalized likelihood (5.2) is equivalent to minimizing

$$\frac{1}{2} \parallel y - X\beta \parallel^2 + n \sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{5.3}$$

Efficient algorithms for finding the minimizer of (5.3) include least angle regression (LARS) for ℓ_1 penalty (Efron et al., 2004), MC+ for quadratic spline penalties (Zhang, 2010), and coordinate-wise optimization algorithm for a general penalty (Friedman et al., 2007).

The generalized linear model may be written, $f(z_i, \theta) = f(z_i, \mu_i)$, where $\mu_i = x_i^T \beta$ and $x_i = (x_{i1}, \ldots, x_{id})^T$. The log-likelihood function, $l(\beta) = \sum_{i=1}^n \log f(z_i, x_i\beta)$, can be locally approximated by

$$\tilde{l}(\beta) = l(\beta^{(0)}) + \nabla l(\beta^{(0)})^T(\beta - \beta^{(0)}) + \frac{1}{2}(\beta - \beta^{(0)})^T \nabla^2 l(\beta^{(0)})(\beta - \beta^{(0)}),$$

where $\beta^{(0)}$ is an initial estimate of β. Let $d(\mu) = \nabla l(\mu)$ and $D(\mu) = -\nabla^2 l(\mu)$, where $l(\mu) = \sum_{i=1}^n \log f(z_i, \mu_i)$. Then $\nabla l(\beta) = X^T d(\mu)$ and $\nabla^2 l(\beta) = -X^T D(\mu) X$. Let

$\mu_0 = X\beta^{(0)}$, $d_0 = d(\mu_0)$, and $D_0 = D(\mu_0)$. Suppose D_0 is positive. Then

$$-\tilde{l}(\beta) = \frac{1}{2} \parallel y^{(1)} - X^{(1)}\beta \parallel^2 + \text{const.},$$

where $X^{(1)} = D_0^{1/2}X$ and $y^{(1)} = D_0^{1/2}X\beta^{(0)} + D_0^{-1/2}d_0$. Thus the penalized maximum likelihood estimator in (5.2) can be approximated by the one-step estimate that minimizes

$$\frac{1}{2} \parallel y^{(1)} - X^{(1)}\beta \parallel^2 + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{5.4}$$

The one-step method was suggested by Zou and Li (2008). Another computationally efficient algorithm for finding the penalized maximum likelihood estimator in generalized linear models is the coordinate-wise optimization algorithm (Friedman et al., 2010).

5.2.4 Model Selection

By model selection we mean the choice of the best model from a set of candidate models that have been produced by penalized maximum likelihood estimation. For each model, we compute the value of the selection criterion. The model having the minimum value of the criterion is selected as the best one. Let $\{\mathcal{S}_k, k = 1, \ldots, K\}$ be the candidate models specified by $\hat{\theta}(\mathcal{S}_k)$. Let $\kappa(\mathcal{S}_k) = k$, and $l_n(\hat{\theta}(\mathcal{S}_k)) \leq l_n(\hat{\theta}(\mathcal{S}_{k+1}))$.

Many criteria have been proposed to select the best model. Some of them may be considered as a special case of the generalized information criterion $\text{GIC} = -2l_n(k) + \alpha k$, where $l_n(k) = l_n(\hat{\theta}(\mathcal{S}_k))$, $k = \kappa(\mathcal{S}_k)$, and $\alpha \geq 0$ is a tuning parameter. We consider a more general form

$$\text{GIC} = -2l_n(k) + \alpha c_k, \tag{5.5}$$

where c_k is a known positive increasing function of the model size k. With $\alpha = 2$ and $c_k = kn/(n-k-1)$, the GIC is the AIC_c (Hurvich and Tsai, 1989). For a linear model, eqn (5.5) reduces to $\text{GIC} = n\log\hat{\sigma}_k^2 + \alpha c_k$, where $\hat{\sigma}_k^2 = \text{RSS}_k/n$, RSS_k is the model residual sum of squares.

Let $\mathcal{S}_{k_n}$ be the selected model and k_n be the model size. Model selection is consistent if $\Pr\{\mathcal{S}_{k_n} = \mathcal{S}_0\} \to 1$ as $n \to \infty$. The model is overfitted if $k_n > k_0$,

and underfitted if $k_n < k_0$. If there is a best model in the set of candidate models, $k_n = k_0$ implies that $\mathcal{S}_{k_n} = \mathcal{S}_0$, and then $\Pr\{\mathcal{S}_{k_n} = \mathcal{S}_0\} = \Pr\{k_n = k_0\}$.

5.3 Asymptotic Properties

Before developing asymptotic properties, we describe some assumptions on the set of candidate models. All limits in this section and the Appendix indicate convergence in probability as $n \to \infty$. For simplicity of notation, let $l_n(k) = l_n(\hat{\theta}(\mathcal{S}_k))$ and $l_n^{(\mathrm{M})}(k) = l_n(\hat{\theta}_{\mathrm{M}}(\mathcal{S}_k))$. And let $\mathcal{S}_{k_0}$ denote the true model. Assume that the following conditions hold:

C1. $E\{|\log f(Z,\theta)|\} < \infty$, and $l_n(1) < \cdots < l_n(K)$.

C2. $\mathcal{S}_{k_0} \in \{\mathcal{S}_k\}$ is identifiable, and $l_n^{(\mathrm{M})}(k_0) - l_n(k_0) = o_p(n)$.

C3. There exists a model $\mathcal{S}_{K_0} \in \{\mathcal{S}_k\}$ satisfying $\mathcal{S}_{k_0} \subset \mathcal{S}_{K_0}$.

From C1 and C2, $\lim n^{-1} l_n(k_0) = \lim n^{-1} l_n^{(\mathrm{M})}(k_0) = E\{\log f(Z,\theta_0)\}$. For $k > k_0$, since $l_n(k) \geq l_n(k_0)$, $\lim n^{-1} l_n(k) = E\{\log f(Z,\theta_0)\}$ and then

$$l_n^{(\mathrm{M})}(k_0) - l_n(k) = o_p(n). \tag{5.6}$$

Let $\Delta_k = 2\{l_n^{(\mathrm{M})}(k) - l_n(k)\}$ represent the bias between the penalized MLE and the MLE. If $\hat{\theta}(\mathcal{S}_k) = \hat{\theta}_{\mathrm{M}}(\mathcal{S}_k)$, $\Delta_k = 0$.

Theorem 5.1. *If $\alpha \leq \varepsilon + (\Delta_{k_0} - \Delta_{K_0})/(c_{K_0} - c_{k_0})$, where ε is a constant, then asymptotically $\Pr\{\kappa(\alpha) = k_0\} \leq q_\varepsilon$ and $\Pr\{\kappa(\alpha) < k_0\} = 0$, where $q_\varepsilon = \Pr\{\chi^2_{K_0 - k_0} \leq \varepsilon(c_{K_0} - c_{k_0})\}$.*

Proofs of all theorems are given in the Appendix. Theorem 5.1 shows how the bias between the penalized MLE and the MLE affects the consistency of the GIC. If $\Delta_{k_0} - \Delta_{K_0} = O_p(n^s)$, $0 < s < 1$, from Theorem 5.1, the GIC with $\alpha \leq n^t$, $t < s$, selects $\mathcal{S}_{k_0}$ with probability tending to zero. In this case, the BIC is not consistent and it selects an overfitted model asymptotically because $\alpha = \log n < n^t$ for a large n.

To characterize the consistency of GIC, we define

$$\gamma = \inf_n \min_{j<k_0} \frac{n^{-1}\{2l_n(k_0) - 2l_n(j)\}}{c_{k_0} - c_j}. \tag{5.7}$$

From

$$\begin{aligned} l_n(k_0) - l_n(j) &\geq l_n(k_0) - l_n^{(\mathrm{M})}(j) \\ &= \{l_n(k_0) - l_n^{(\mathrm{M})}(k_0)\} + \{l_n^{(\mathrm{M})}(k_0) - l_n^{(\mathrm{M})}(j)\}, \end{aligned}$$

and (5.1) and (5.6), we have $\gamma \geq 2\Delta/(c_{k_0} - c_1)$.

Theorem 5.2. *Assume* $\Delta_{k_0} = O_p(1)$ *and* $\Delta_{K_0} = O_p(1)$. *Let* $\mathrm{D}_K = 2l_n^{(\mathrm{M})}(K) - 2l_n^{(\mathrm{M})}(k_0)$. *Let* $\alpha = \alpha_n$ *and* $n^{-1}\alpha_n \to r$. *Then asymptotically:*
(i) If $\alpha_n < \infty$, GIC *selects either the true model or an overfitted model.*
(ii) If $\alpha_n \geq (\mathrm{D}_K + \Delta_{k_0})/(c_{k_0+1} - c_{k_0})$ *and* $r < \gamma$, GIC *is consistent.*
(iii) If $\alpha_n \to \infty$ *and* $r > \gamma$, GIC *selects an underfitted model.*

With suitable penalty, the penalized likelihood methods asymptotically produce a set of models in which there is a best model having the oracle property. Then $l_n(k_0) = l_n^{(\mathrm{M})}(k_0)$ and $\Delta_{k_0} = 0$ with probability tending to one. Moreover, as the regularization parameter is near to zero, the penalized MLE reduces to the MLE, and then $\Delta_{K_0} = 0$. The asymptotic properties of the GIC with the penalized MLE are summarized in Corollary 5.1.

Corollary 5.1. *Let* $\{\mathcal{S}_k\}$ *be produced by the penalized* MLE *having the oracle property. Let* $\mathrm{D}_K = 2l_n^{(\mathrm{M})}(K) - 2l_n^{(\mathrm{M})}(k_0)$. *Let* $\alpha = \alpha_n$ *and* $n^{-1}\alpha_n \to r$. *Then asymptotically:*
(i) If $\alpha_n < \infty$, GIC *selects either the true model or an overfitted model.*
(ii) If $\alpha_n \geq \mathrm{D}_K/(c_{k_0+1} - c_{k_0})$ *and* $r < \gamma$, GIC *is consistent.*
(iii) If $\alpha_n \to \infty$ *and* $r > \gamma$, GIC *selects an underfitted model.*

This Corollary follows directly from Theorem 5.2. Corollary 5.1 (i) is the same as that given in Zhang et al. (2010). If $\mathrm{D}_K = \infty$, the largest model, $\mathcal{S}_K$, is always selected. In this case, GIC can not work. This usually takes place in the case where $n < K$. We may avoid this case by reducing the dimension. If K is finite, $\mathrm{D}_K = l_n^{(\mathrm{M})}(K) - l_n^{(\mathrm{M})}(k_0) \sim \chi^2_{K-k_0}$ asymptotically. Then the consistent condition in Corollary 5.1 (ii) is that $\alpha_n \to \infty$ and $r < \gamma$. So the following Corollary also holds.

Corollary 5.2. *Let* $\{\mathcal{S}_k\}$ *be produced by the* MLE *and* K *be finite. Let* $\alpha = \alpha_n$ *and* $n^{-1}\alpha_n \to r$. *Then asymptotically:*
(i) If $\alpha_n < \infty$, GIC *selects either the true model or an overfitted model.*
(ii) If $\alpha_n \to \infty$ *and* $r < \gamma$, GIC *is consistent.*
(iii) If $\alpha_n \to \infty$ *and* $r > \gamma$, GIC *selects an underfitted model.*

Under the conditions in Theorem 5.2 or Corollary 5.1, the GIC with α_n bounded is a AIC-type criterion having the same asymptotic property as AIC, while the GIC with α_n unbounded and $\lim n^{-1}\alpha_n < \gamma$, is a BIC-type criterion having the same asymptotic property as BIC.

Suppose that $\hat{\theta}(\mathcal{S}_j) \to \theta^*(\mathcal{S}_j)$, $\hat{\theta}(\mathcal{S}_{k_0}) \to \theta_0$, and $l_n^{(\mathrm{M})}(j) - l_n(j) = o_p(n)$ for $1 \le j \le K$. Then $n^{-1}\{l_n(k_0) - l_n(j)\} \to E\{\log f(Z,\theta_0) - \log f(Z,\theta^*(\mathcal{S}_j))\}$, and

$$\gamma = \min_{j<k_0} \frac{2E\{\log f(Z,\theta_0) - \log f(Z,\theta^*(\mathcal{S}_j))\}}{c_{k_0} - c_j}. \tag{5.8}$$

For the linear model $Y = x^T\beta + \sigma\epsilon$, where ϵ is standard normal and σ^2 is the error variance, $2\log f(Z,\beta) = -(Y - x^T\beta)^2/\sigma^2 - \log \pi\sigma^2$. Suppose that $\hat{\beta}(\mathcal{S}_j) \to \beta^*(\mathcal{S}_j)$ and $\hat{\beta}(\mathcal{S}_{k_0}) \to \beta_0$. Let $c_k = k$. From (5.8),

$$\gamma = \min_{j<k_0} \frac{E\{|x^T(\beta_0 - \beta^*(\mathcal{S}_j))|^2\}}{(k_0 - j)\sigma^2}. \tag{5.9}$$

So γ may be viewed as an average signal-to-noise ratio.

5.4 Numerical Illustration

In this section we present simulation examples to numerically examine the finite-sample properties of GIC for model selection with the penalized MLE using AIC and BIC. When the penalized MLE has the oracle property, the AIC-type and BIC-type share the asymptotic properties of AIC and BIC, respectively. Two linear regressions and a logistic model are examined. For linear models, the R package *plus* (Zhang, 2010) is used. For the logistic model, the one-step estimates are used (Zou and Li, 2008). A large number of simulations, 10^4, was used for each case, so the error of estimation is negligible. The captions below each figure give the 95% margin of error.

5.4.1 Linear Regression Models

We consider the linear models with regression coefficients given by the d-dimensional vector β and specified by $Y = x^T\beta + \sigma\epsilon$, where ϵ is standard normal, σ^2 is the error variance, x is d-variate normal with covariance matrix $(0.5^{|i-j|})_{d\times d}$.

In the first example we take $d = 8$, $k_0 = 3$, $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\sigma = 1$. The components of X and ε are standard normal. Sample sizes are $n = 20$, 60, and 100. The AIC and BIC are used to select the model and the candidate models are produced by best subset, LASSO, SCAD, and MCP. It was found that the proportion of underfit models was zero or very close to zero in each case. The proportion of correctly selected models is shown Figure 5.1. The best subset method, SUBSET, outperforms LASSO but not SCAD and MCP. The BIC, SUBSET, SCAD and MCP improve markedly as n increase with SCAD and MCP slightly better than SUBSET. As shown in Theorem 5.1, AIC is overfitting and this behaviour is confirmed in Figure 5.1.

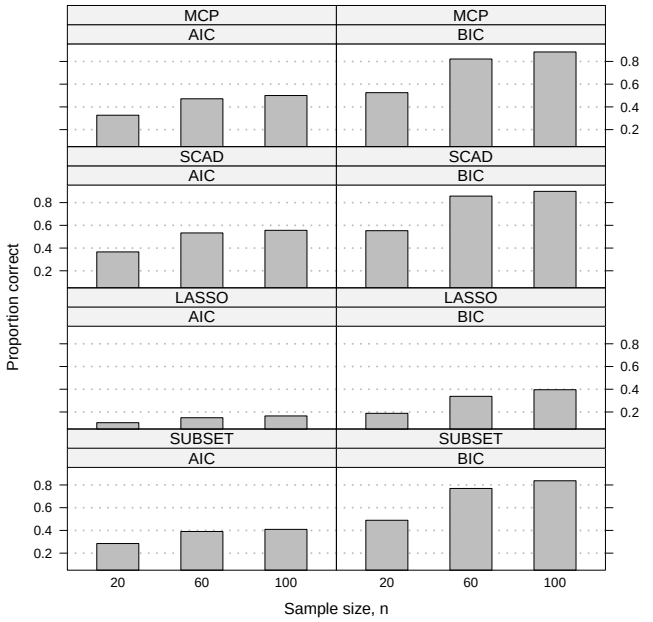

Figure 5.1: Proportion of models correctly selected in linear regression example with $d = 8$. The 95% margin of error is less than 0.01.

The conditional prediction error, PE $= E\{(Y - x^T(\mathcal{S})\hat{\beta}(\mathcal{S}))^2\}$, was computed using 10^5 independent test samples for each simulation. The average of all 10^4 simulations, estimates the prediction error that is shown in Figure 5.2 for the various methods. The average prediction error when LASSO is used for variable selection only and the usual least-square estimates are then used for the prediction is also shown in Figure 5.2 for the method denoted by LASSOLS. We see that LASSOLS outperforms LASSO and SUBSET as expected but it is not quite as good as SCAD and MCP. The ORACLE prediction error is obtained assuming the correct variables are known a priori

and then the usual least squares estimates of the parameters are used. When $n = 100$, SCAD and MCP are close to the ORACLE prediction error when the BIC is used. As might be expected, due to overfitting, the AIC performance is generally much poorer than with the BIC.

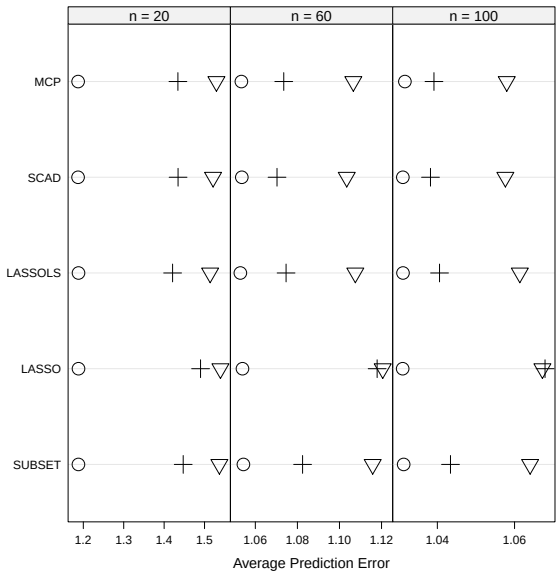

Figure 5.2: Average prediction errors in linear regression with $d = 8$. ORACLE: $\circ$, AIC: ∇, BIC: $+$. 95% MOE $< 0.01, 0.002, 0.001$ for $n = 20, 60, 100$ respectively.

The performance of AIC_α in the large feature space is examined using $d = 50$, $n = 100$, $k_0 = 7$, $\beta_0 = (1, -0.5, 0.7, -1.2, -0.9, 0.3, 0.55)$ and $\sigma = 0.3$. The GIC with $\alpha = 2, 4, \ldots, 20$ are used to select the best model from the candidate models produced by LASSO, SCAD, and MCP. The proportion of underfit, correct and overfit models are shown in Figure 5.3. As α increases, the number of the overfitted models decreases, and the number of the underfitted models increases. The BIC corresponds to $\alpha = 4.6$. Using SCAD or MCP with GIC for $\alpha \geq 4$ produces a high proportion of correct models but LASSO always produces many overfit models even for large values of α. The underfitting condition in Corollary 5.1 is illustrated by the fact that when $\alpha = 20$ all penalty methods produce some underfit models.

5.4.2 Logistic Regression Model

In this example, the data are generated from the logistic regression model with $d = 25$, $k_0 = 5$ and $\beta_0 = (2.5, -1.9, 2.8, -2.2, 3)$ for samples of size, $n = 200, 250, 300$.

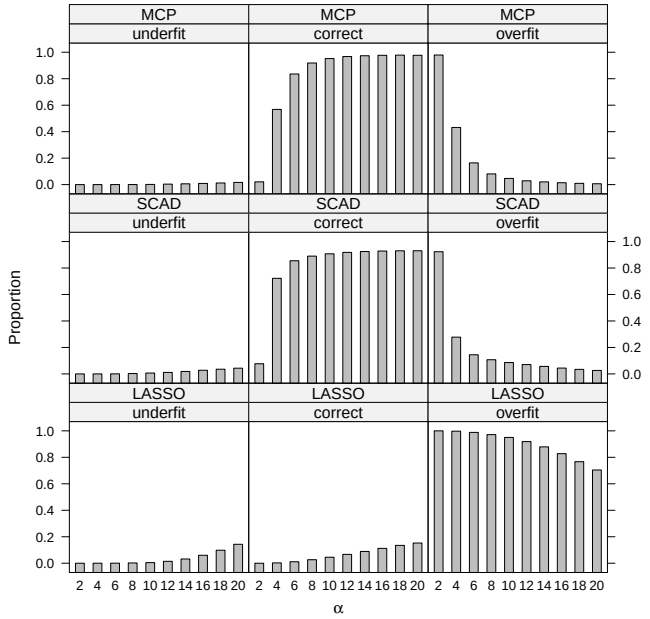

Figure 5.3: Proportions of underfitted, correctly fitted and overfitted models in linear regression example with $d = 50$. The 95% margin of error is less than 0.01.

As shown in Corollary 5.1, for the penalized likelihood methods having the oracle property, the BIC-type criteria are consistent but the AIC-type are not and this is in agreement with the results shown in Figure 5.4. For SCAD and MCP, the number of the correctly fitted models by BIC increases as n becomes larger but there is only a very slight increase with the AIC. For LASSO, both AIC and BIC select a small number of correctly fitted models. For the penalized likelihood method such as LASSO having no oracle property, the BIC is not consistent.

5.5 Conclusions

Usually there is a bias between the penalized MLE and the MLE. We analyzed how the bias affects the consistency of GIC, and derived the conditions under which the GIC is overfitting, consistent, or underfitting. If the penalized MLE has the oracle property, the GIC may be categorized into three types: AIC-type, BIC-type, and others that are asymptotically underfitting, and the AIC-type and BIC-type have the same asymptotic properties as AIC and BIC, respectively. If the penalized MLE does not possess the oracle property, as is the case with LASSO, the BIC-type may not be consistent.

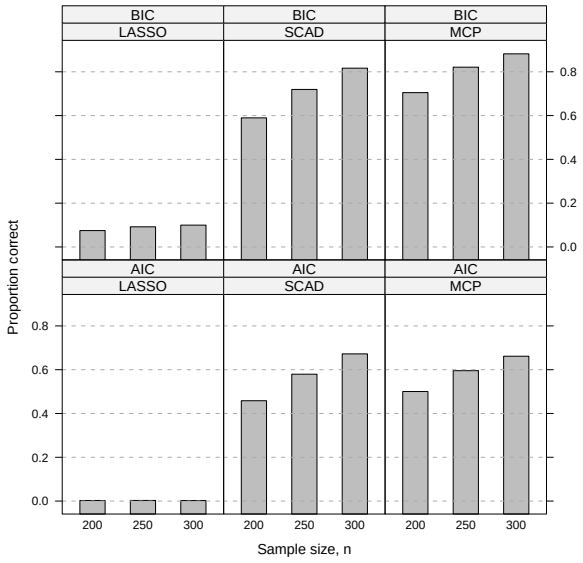

Figure 5.4: Percentages of correctly fitted models in logistic regression with $d = 25$. The 95% margin of error is less than 0.01.

5.6 Appendix

5.6.1 Lemmas

The following lemmas are useful for the proofs of theorems.

Lemma 5.1. GIC *can select the model* $\mathcal{S}_k$ *if and only if*

$$\max_{j>k} A_{k,j} \leq \alpha \leq \min_{j<k} A_{k,j},$$

where $A_{k,j} = 2\{l_n(j) - l_n(k)\}/(c_j - c_k)$ *for* $j \neq k$, $k = 1, \ldots, K$.

Proof. GIC selects model $\mathcal{S}_k$ if and only if for $j \neq k$,

$$-2\{l_n(j) - l_n(k)\} + \alpha(c_j - c_k) \geq 0,$$

that is,

$$\max_{j>k} A_{k,j} \leq \alpha \leq \min_{j<k} A_{k,j}.$$

□

Define $\min_{j<1} A_{1,j} = \infty$ and $\max_{j>K} A_{K,j} = 0$. Let $\kappa(\alpha)$ be the model size selected by GIC. The probability of selecting the model $\mathcal{S}_k$ is

$$\Pr\{\kappa(\alpha) = k\} = \Pr\{\max_{j>k} A_{k,j} \leq \alpha \leq \min_{j<k} A_{k,j}\}. \tag{5.10}$$

Lemma 5.2. *Let $\alpha = \alpha_n$ and $n^{-1}\alpha_n \to r$. Let $\mathrm{D}_K = 2l_n^{(\mathrm{M})}(K) - 2l_n^{(\mathrm{M})}(k_0)$. If $\alpha_n \geq (\mathrm{D}_K + \Delta_{k_0})/(c_{k_0+1} - c_{k_0})$ and $r < \gamma$ then $\Pr\{\kappa(\alpha) = k_0\} \to 1$.*

Proof. Let $r < \gamma$. From (5.7), $n^{-1}\alpha_n < \min_{j<k_0} n^{-1}A_{k_0,j}$ for n large enough. From (5.10)

$$\begin{aligned}
\Pr\{\kappa(\alpha) = k_0\} &= \Pr\{\max_{j>k_0} A_{k_0,j} \leq \alpha_n, \alpha_n \leq \min_{j<k_0} A_{k_0,j}\} \\
&= \Pr\{\max_{j>k_0} A_{k_0,j} \leq \alpha_n, n^{-1}\alpha_n \leq \min_{j<k_0} n^{-1}A_{k_0,j}\} \\
&= \Pr\{\max_{j>k_0} A_{k_0,j} \leq \alpha_n\} \\
&= \Pr\{\max_{j>k_0} [2l_n(j) - 2l_n(k_0)]/(c_j - c_{k_0}) \leq \alpha_n\} \\
&\geq \Pr\{\max_{j>k_0} [2l_n^{(\mathrm{M})}(K) - 2l_n(k_0)]/(c_j - c_{k_0}) \leq \alpha_n\} \\
&= \Pr\{2[l_n^{(\mathrm{M})}(K) - l_n(k_0)] \leq \alpha_n(c_{k_0+1} - c_{k_0})\} \\
&= \Pr\{\mathrm{D}_K + \Delta_{k_0} \leq \alpha_n(c_{k_0+1} - c_{k_0})\} \\
&= 1.
\end{aligned}$$

□

Lemma 5.3. *Let $\alpha = \alpha_n$ and $n^{-1}\alpha_n \to r$. If $r \geq \gamma$ then $\Pr\{\kappa(\alpha) = k_0\} \leq \delta_{r,\gamma}$ and $\Pr\{\kappa(\alpha) > k_0\} \to 0$, where $\delta_{r,\gamma}$ is the Kronecker delta.*

Proof. Let $r \geq \gamma$, and n be large enough. Then there exists a constant ε such that $n^{-1}\alpha_n \geq \gamma - \varepsilon > 0$. For $j > k_0$, from (5.6), $n^{-1}\{l_n(j) - l_n^{(\mathrm{M})}(k_0)\} = o_p(1)$. Thus $j > k_0$,

$$A_{k_0,j} = \frac{2\{l_n(j) - l_n^{(\mathrm{M})}(k_0)\} + 2\{l_n^{(\mathrm{M})}(k_0) - l_n(k_0)\}}{c_j - c_{k_n}} < \alpha_n.$$

Let $k > k_0$. From (5.10)

$$\Pr\{\kappa(\alpha) = k\} \leq \Pr\{\alpha_n \leq \min_{j<k} A_{k,j}\} \leq \Pr\{\alpha_n \leq A_{k_0,k}\} = 0.$$

From (5.10)

$$\begin{aligned}
\Pr\{\kappa(\alpha) = k_0\} &= \Pr\{\max_{j>k_0} A_{k_0,j} \leq \alpha_n, \alpha_n \leq \min_{j<k_0} A_{k_0,j}\} \\
&= \Pr\{\alpha_n \leq \min_{j<k_0} A_{k_0,j}\} \\
&= \Pr\{n^{-1}\alpha_n \leq \min_{j<k_0} n^{-1} A_{k_0,j}\} \\
&\leq \Pr\{r \leq \gamma\} \\
&= \delta_{r,\gamma}.
\end{aligned}$$

□

5.6.2 Proofs of Theorems

Theorem 5.1. From (5.10),

$$\begin{aligned}
\Pr\{\kappa(\alpha) = k_0\} &\leq \Pr\{\max_{j>k_0} \frac{2l_n(j) - 2l_n(k_0)}{c_j - c_{k_0}} \leq \alpha\} \\
&\leq \Pr\{\frac{2l_n(K_0) - 2l_n(k_0)}{c_{K_0} - c_{k_0}} \leq \alpha\}.
\end{aligned}$$

Since

$$\begin{aligned}
&l_n(K_0) - l_n(k_0) \\
&= \{l_n^{(\mathrm{M})}(K_0) - l_n^{(\mathrm{M})}(k_0)\} + \{l_n^{(\mathrm{M})}(k_0) - l_n(k_0)\} - \{l_n^{(\mathrm{M})}(K_0) - l_n(K_0)\} \\
&= \{l_n^{(\mathrm{M})}(K_0) - l_n^{(\mathrm{M})}(k_0)\} + (\Delta_{k_0} - \Delta_{K_0})/2,
\end{aligned}$$

it holds asymptotically that

$$\begin{aligned}
\Pr\{\kappa(\alpha) = k_0\} &\leq \Pr\{2[l_n^{(\mathrm{M})}(K_0) - l_n^{(\mathrm{M})}(k_0)] \leq \alpha(c_{K_0} - c_{k_0}) - (\Delta_{k_0} - \Delta_{K_0})\} \\
&\leq \Pr\{2[l_n^{(\mathrm{M})}(K_0) - l_n^{(\mathrm{M})}(k_0)] \leq \varepsilon(c_{K_0} - c_{k_0})\} \\
&= \Pr\{\chi^2_{K_0-k_0} \leq \varepsilon(c_{K_0} - c_{k_0})\} \\
&= q_\varepsilon.
\end{aligned}$$

Let $k < k_0$. From (5.10),

$$\begin{aligned}
\Pr\{\kappa(\alpha) = k\} &\leq \Pr\{\max_{j>k} \frac{2l_n(j) - 2l_n(k)}{c_j - c_k} \leq \alpha\} \\
&\leq \Pr\{\frac{2l_n(k_0) - 2l_n(k)}{c_{k_0} - c_k} \leq \alpha\} \\
&\leq \Pr\{\frac{2l_n(k_0) - 2l_n^{(\mathrm{M})}(k)}{c_{k_0} - c_k} \leq \alpha\} \\
&= \Pr\{\frac{2n^{-1}[l_n(k_0) - l_n^{(\mathrm{M})}(k)]}{c_{k_0} - c_k} \leq n^{-1}\alpha\}.
\end{aligned}$$

From (5.1) and (5.6),

$$\begin{aligned}
n^{-1}\{l_n(k_0) - l_n^{(\mathrm{M})}(k)\} &= n^{-1}\{l_n^{(\mathrm{M})}(k_0) - l_n^{(\mathrm{M})}(k)\} - n^{-1}\{l_n^{(\mathrm{M})}(k_0) - l_n(k_0)\} \\
&\geq \Delta - o_p(1).
\end{aligned}$$

Since $n^{-1}\alpha \to 0$, $\Pr\{\kappa(\alpha) = k\} = 0$. □

Theorem 5.2. It follows directly from Theorem 5.1 and Lemmas 5.2 and 5.3. □

Chapter 6

SUMMARIES AND FUTURE DIRECTIONS

This thesis primarily studied the properties and improvement of the information criteria including the generalized information criterion and the family of Bayesian information criteria. The following issues have been addressed: i) non-asymptotic and asymptotic properties of GIC with LS or MLE algorithms; ii) asymptotic properties of GIC with penalized MLE algorithms for high dimensional space; and iii) improved information criteria for model selection and choice of tuning parameter.

The non-asymptotic properties include the probability of selecting a model and its computation, and the interval constraints under which a specified model can be selected. The asymptotic properties provide the conditions under which the information criteria are overfitting, consistent, or underfitting. Three procedures for improving the GIC have been proposed. Two of them are the adaptive GIC that is based on the probability of selecting a model. The other is the GIC with controlling the overfitting level.

There are a lot of issues that are not covered in this thesis. Some topics for possible future research are indicated in the following sections.

Model Dimension Reduction

When dimensionality can be reduced, model selection may be more accurate. The computational burden can be reduced dramatically.

Missing Data

We need to develop selection procedures that more effectively deal with missing values instead of simply discarding them. Missing values should automatically be imputed or estimated at each stage in the selection process.

Multivariate Models

This thesis dealt only with response or output variables that are univariate. The extension of this thesis to deal with the more general case of multivariate outputs is straightforward but the details and algorithmic implementations would be considerably more complex.

Biomedical Applications

There are many applications of the methods discussed in this thesis in the medical and biomedical area. Model selection has attracted substantial attention in high-dimensional genomic and proteomic data analysis. Model selection is extensively used for survival and longitudinal data analysis and identification of the quantitative trait loci. A growing important area of medical research deals with the assessment of a drug or the effect of a new medical procedure by observing the actual historical outcomes. The data thus comprise a time series and an intervention analysis model may be used to assess and describe the overall effect. Model selection is important in selecting model components describing the errors, the covariates, if any, and the intervention itself. A very interesting and important example is discussed by (Juurlink et al., 2004) who show that deaths in the target population increased after a new drug Aldactone was introduced. This drug had been extensively tested in clinical trials but proved a disaster in actual clinical usage. This disaster was detected and its magnitude quantified using time series intervention analysis.

REFERENCES

Akaike, H., 1969. Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics 21 (1), 243–247.

Akaike, H., 1970. Statistical predictor identification. Annals of the Institute of Statistical Mathematics 22 (1), 203–217.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716–723.

Akaike, H., 1979. A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. Biometrika 66 (2), 237–242.

Atkinson, A. C., 1980. A note on the generalized information criterion for choice of a model. Biometrika 67 (2), 413–418.

Bera, A. K., 2000. Hypothesis testing in the 20th century with a special reference to testing with misspecified models. Statistics for the 21st Century: Methodologies for Applications of the Future, C.R. Rao and Gabor J. Szekely, Editors, Marcel Dekkar, 2000, 33–92.

Bhansali, R. J., 1986. A derivation of the information criteria for selecting autoregressive models. Advances in Applied Probability 18 (2), 360–387.

Bhansali, R. J., Downham, D. Y., 1977. Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion. Biometrika 64 (3), 547–551.

Box, G., Jenkins, G. M., Reinsel, G. C., 2005. Time Series Analysis: Forecasting & Control (4th Edition). Wiley, New York.

Breiman, L., 1996. Heuristics of instability and stabilization in model selection. The Annals of Statistics 24, 2350–2383.

Breiman, L., 2001. Statistical modeling: The two cultures. Statistical Science 16 (3), 199–231.

Burnham, K. P., Anderson, D. R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd ed. Springer-Verlag, New York.

Chatfield, C., 1979. Inverse autocorrelations. Journal of the Royal Statistical Society A 142 (3), 363–377.

Chen, J., Chen, Z., 2008. Extended Bayesian information criterion for model selection with large model spaces. Biometrika 95 (3), 759–771.

Cleveland, W. S., 1972. The inverse autocorrelations of a time series and their applications. Technometrics 14 (2), 277–293.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression (with discussion). The Annals of Statistics 32 (2), 407–499.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Fan, J., Lv, J., 2010. A selective overview of variable selection in high dimensional feature space. Statistica Sinica 20 (1), 101–148.

Foster, D. P., George, E. I., 1994. The risk inflation criterion for multiple regression. The Annals of Statistics 22, 1947–1975.

Frank, I. E., Friedman, J. H., 1993. A statistical view of some chemometrics regression tools (with discussion). Technometrics 35, 109–148.

Friedman, J., Hastie, T., Hoefling, H., Tibshirani, R., 2007. Pathwise coordinate optimization. Annals of Applied Statistics 2 (1), 302–332.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 (1), 1–22.

Furnival, G., Wilson, R., 1974. Regression by leaps and bounds. Technometrics 16 (16), 499–511.

Gatu, C., 2006. Branch-and-bound algorithms for computing the best-subset regression models. Journal of Computational and Graphical Statistics 15, 139–156.

George, E. I., Foster, D. P., 2000. Calibration and empirical Bayes variable selection. Biometrika 87 (4), 731–747.

Granger, C., Jeon, Y., 2004. Forecasting performance of information criteria with many macro series. Journal of Applied Statistics 31 (10), 1227–1240.

Hansen, M. H., Yu, B., 2001. Model selection and the principle of minimum description length. Journal of the American Statistical Association 96 (454), 746–774.

Hastie, T., Tibshirani, R., Friedman, J. H., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd Edition. Springer-Verlag, New York.

Hipel, K. W., McLeod, A. I., 1994. Time Series Modelling of Water Resources and Environmental Systems. Elsevier, Amsterdam.

Hofmann, M., Gatu, C., Kontoghiorghes, E. J., 2007. Efficient algorithms for computing the best subset regression models for large-scale problems. Computational Statistics & Data Analysis 52 (1), 16–29.

Hurvich, C. M., Tsai, C.-L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Juurlink, D. N., Mamdani, M. M., Lee, D. S., Alexander Kopp, P. C. A., Laupacis, A., Redelmeier, D. A., 2004. Rates of hyperkalemia after publication of the randomized aldactone evaluation study. New England Journal of Medicine 351, 543–551.

Kay, S. M., 1988. Modern Spectral Estimation. Signal Processing. Prentice Hall, Englewood Cliffs, New Jersey.

Kubokawa, T., Robert, C. P., Saleh, A. K. M. E., 1993. Estimation of noncentrality parameters. The Canadian Journal of Statistics 21 (1), 45–57.

Lahiri, P., 2001. Model Selection. Edited by P. Lahiri. Beachwood, Ohio: Institute of Mathematical Statistics.

Linhart, H., Zucchini, W., 1986. Model Selection. John Wiley & Sons, New York.

Mallows, C. L., 1973. Some comments on C_p. Technometrics 15, 661–675.

McLeod, A. I., 1993. Parsimony, model adequacy and periodic correlation in forecasting time series. International Statistical Review 61 (3), 387–393.

McLeod, A. I., Xu, C., 2010. bestglm: Best Subset GLM.
URL http://CRAN.R-project.org/package=bestglm

McLeod, A. I., Zhang, Y., Xu, C., 2010. FitAR: Subset AR Model Fitting.
URL http://CRAN.R-project.org/package=FitAR

McQuarrie, A. D. R., Tsai, C. L., 1998. Regression and Time Series Model Selection. World Scientific Publishing Company, Singapore.

Miller, A., 2002. Subset Selection in Regression. Chapman and Hall, New York.

Myers, R., 1990. Classical and Modern Regression with Applications. The Duxbury Advanced Series in Statistics and Decision Sciences. PWS-KENT Publishing Company, Boston.

Nishii, R., 1984. Asymptotic properties of criteria for selection of variables in multiple regression. The Annals of Statistics 12, 758–765.

Percival, D. B., Walden, A. T., 1993. Spectral Analysis for Physical Applications. Cambridge University Press.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org

Rao, C. R., 1973. Linear Statistical Inference and Its Applications, 2nd Edition. John Wiley & Sons.

Rao, J. S., 1999. Bootstrap choice of cost complexity parameter for better subset selection. Statistica Sinica 9, 273–287.

Rissanen, J., 1978. Modeling by shortest data description. Automatica 14, 465–471.

Rissanen, J., 1983. A universal prior for integers and estimation by minimum description length. The Annals of Statistics 11 (2), 416–431.

Rissanen, J., 2007. Information and Complexity in Statistical Modeling. Springer-Verlag, New York.

Robert, C. P., 2007. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, 2nd Edition. Springer-Verlag, New York.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6 (2), 461–464.

Shao, J., 1993. Linear model selection by cross-validation. Journal of the American Statistical Association 88, 486–494.

Shao, J., 1996. Bootstrap model selection. Journal of the American Statistical Association 91, 655–665.

Shao, J., 1997. An asymptotic theory for linear model selection. Statistica Sinica 7 (2), 221–262.

Shao, J., Rao, J. S., 2000. The GIC for model selection: a hypothesis testing approach. Journal of Statistical Planning and Inference 88, 215–231.

Shen, X., Ye, J., 2002. Adaptive model selection. Journal of the American Statistical Association 97, 210–221.

Shibata, R., 1984. Approximate efficiency of a selection procedure for the number of regression variables. Biometrika 71, 43–49.

Sin, C.-Y., White, H., 1996. Information criteria for selecting possibly misspecified parametric models. Journal of Econometrics 71, 207–225.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

Tibshirani, R., Knight, K., 1999. The covariance ination criterion for adaptive model selection. Journal of the Royal Statistical Society, Series B 61, 529–546.

Tong, H., 1977. Some comments on the Canadian lynx data. Journal of the Royal Statistical Society A 140 (4), 432–436.

Vapnik, V. N., 2000. The nature of statistical learning theory, 2nd Edition. Springer-Verlag, New York.

Wang, H., Li, R., Tsai, C.-L., 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94, 553–568.

Yang, Y., 2005. Can the strengths of AIC and BIC be shared? Biometrika 92 (4), 937–950.

Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics 38 (2), 894–942.

Zhang, Y., 2009. Model selection: A Lagrange optimization approach. Journal of Statistical Planning and Inference 139, 3142–3159.

Zhang, Y., Li, R., Tsai, C.-L., 2010. Regularization parameter selections via generalized information criterion. Journal of the American Statistical Association 105 (489), 312–323.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 1418–1429.

Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models (with discussion). The Annals of Statistics 36 (4), 1509–1566.

CURRICULUM VITAE

Changjiang Xu

Post-secondary Education and Degrees

- **Ph.D.** (*Statistics*), the University of Western Ontario, London, Canada, 2010.
- **Ph.D.** (*Control Theory*), Southeast University, China, 1997.
- **M.Sc.** (*Applied Mathematics*), Zhejiang University, China, 1989.
- **B.Sc.** (*Mathematics*), Nanjing University, China, 1986.

Honours and Awards

- Western Graduate Research Scholarship, 2007 – 2010
- Graduate Thesis Research Award, 2009
- Robert and Ruth Lumsden Graduate Fellowship, 2009

Presentation

1. "Properties of Linear Model Selection", poster presentation at *Joint Statistics Meetings*, Vancouver, August 4, 2010.
2. "Adjustable Bayesian Information Criterion for Model Selection", oral presentation at *Annual Meeting of the Statistical Society of Canada*, SSC 2009, Vancouver, May 2009.

Articles for Journal Publications

1. Xu, Changjiang and McLeod, A.I. (2010). "Asymptotic Properties of Generalized Information Criterion for Choosing the Regularization Parameter", *submitted*.
2. Xu, Changjiang and McLeod, A.I. (2010). "Bayesian Information Criterion with Bernoulli Prior", *submitted*.
3. Xu, Changjiang and McLeod, A.I. (2010). "Model Selection Using Generalized Information Criterion", *Submitted*.